information

Review

Feature Encodings and Poolings for Action and Event Recognition: A Comprehensive Survey

Changyu Liu [1,2,*], Qian Zhang [1], Bin Lu [3,*] and Cong Li [4,5]

1 College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China; janetcizhang@gmail.com
2 School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
3 School of Computer Science, Wuyi University, Jiangmen 529020, China
4 College of Computer Science, Sichuan Normal University, Chengdu 610068, China; jkxy_lc@sicnu.edu.cn
5 School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 611731, China
* Correspondence: yezhich@gmail.com (C.L.); lbscut@gmail.com (B.L.); Tel.: +86-20-8528-5396 (C.L.)

Received: 23 August 2017; Accepted: 24 October 2017; Published: 29 October 2017

Abstract: Action and event recognition in multimedia collections is relevant to progress in cross-disciplinary research areas including computer vision, computational optimization, statistical learning, and nonlinear dynamics. Over the past two decades, action and event recognition has evolved from earlier intervening strategies under controlled environments to recent automatic solutions under dynamic environments, resulting in an imperative requirement to effectively organize spatiotemporal deep features. Consequently, resorting to feature encodings and poolings for action and event recognition in complex multimedia collections is an inevitable trend. The purpose of this paper is to offer a comprehensive survey on the most popular feature encoding and pooling approaches in action and event recognition in recent years by summarizing systematically both underlying theoretical principles and original experimental conclusions of those approaches based on an approach-based taxonomy, so as to provide impetus for future relevant studies.

Keywords: action and event recognition; multimedia collections; computer vision; feature encodings and poolings

1. Introduction

More and more research efforts within the computer vision community have focused on recognizing actions and events from uncontrolled videos over the past two decades. There are many promising applications for action and event recognition, such as abnormal action and event recognition in surveillance applications [1–4], interaction action and event recognition in entertainment applications [5–8], and home-based rehabilitation action and event recognition in healthcare applications [9–12], and many other analogous applications such as in [13–18]. According to the definition given by NIST [19], an event is a complex activity occurring at a specific place and time, which involves people interacting with other people and/or objects, and consists of a number of human actions, processes, and activities. Feature representation approaches, pattern recognition models and performance evaluation strategies are three key components in action and event recognition [20]. Compared to the other two key components, feature representation approaches, which should contain robust appearance and motion information, played a more critical role in video analysis. There are three key components in feature representation approaches, which are feature extractions, feature encodings, and feature poolings.

Feature extractions care mainly about how to extract required features from specified multimedia collections. On the one hand, there are many approaches proposed for feature extraction using specified sensors in previous research work. For example, Lara et al. [21] surveyed the state of the art in wearable sensors-based human activity recognition, and categorized feature extraction from time series data into statistical approaches and structural approaches, with measured attributes of acceleration, environmental signals and vital signs. Jalal et al. [22] developed a life logging system, where both magnitude features and direction angle features were extracted from depth silhouettes of human activities captured by a depth camera. Yang et al. [23] applied a low-bandwidth wearable motion sensor network to recognize human actions distributed on individual sensor nodes and a base station computer, based on a set of LDA features. Song et al. [24] proposed a robust feature approach, namely the body surface context, to encode the cylindrical angular of the difference vector according to the characteristics of human body, for action recognition from videos of depth camera. Jalal et al. [25] presented a methodology for human activity recognition-based smart home application, by extraction of both magnitude features and direction features from human silhouettes in a depth camera. Althloothi et al. [26] extracted two sets of 3D spatiotemporal features via a Kinect sensor for human activity recognition, where the shape features were from the surface points using spherical harmonics coefficients, and the motion features were from the end points of the distal limb segments. Besides, similar research work in the area can be found in references [27–29].

On the other hand, there are also a number of approaches proposed for extraction of multimedia features in previous research work, including audio features, visual features, text features, and hybrid features. For example, Li et al. [30] proposed the extraction of deep audio features for acoustic event detection by a multi-stream hierarchical deep neural network. Kumar et al. [31] proposed a unified approach that adopted strongly and weakly labeled data for audio event and scene recognition based on a mel-ceptra coefficients feature. Farooq et al. [32] constructed a feature-structured framework by using skin joints features and self-organizing map for 3D human activity detection, tracking and recognition from RGB-D video sequences. Siswanto et al. [33] verified by experiments that PCA-based Eigenface outperformed LDA-based Fisherface for facial recognition for biometrics-based time attendance purposes. Manwatkar et al. [34] designed an automatic image-text recognition system by using matrix features and Kohonen neural networks. Chang et al. [35] invented a source-concept bi-level semantic representation analysis framework for multimedia event detection. Jalal et al. [36] proposed a hybrid feature representation approach called depth silhouettes context, which is fused of invariant features, depth sequential silhouettes features and spatiotemporal body joints features, for human activity recognition based on embedded Hidden Markov Models. Kamal et al. [37] presented a framework for 3D human body detection, tracking and recognition from depth video sequences using spatiotemporal features and modified HMM. Jalal et al. [38] proposed a hybrid multi-fused spatiotemporal feature representation approach that concatenated four skeleton joint features and one body shape feature to recognize human activity from depth video.

The feature encodings and poolings concern primarily how to organize effectively the extracted features, so that overall performances in action and event recognition, such as recognition precision, recognition robustness and computation complexity, can be further improved. The difference between action representation and event representation lies mainly in their feature extraction approaches, whereas these two representations usually share same feature encodings and poolings. Due to complex spatiotemporal relationships between actions and events in multimedia collections, feature encodings and poolings are becoming increasingly important in feature representation approaches. There have been many survey papers available on feature representation approaches of action and event recognition [39–48]. However, most of them focused on reviewing proposed feature extraction approaches, and thus did not take into comprehensive account feature encoding and pooling approaches, which have been widely used in both image and video analysis. The purpose of the paper is to conduct a complete survey on the most popular feature encoding and pooling approaches in action and event recognition from recent years.

2. Feature Encoding and Pooling Taxonomy

The hierarchical taxonomy of the survey paper is shown in Figure 1. These proposed methodologies of feature encodings and poolings for action and event recognition from recent years can be classified into four categories, which are 2D encodings, 3D encodings, general poolings, and particular poolings, where most of the methodologies the paper surveyed are from papers of either top conferences or top journals in computer vision and pattern recognition.

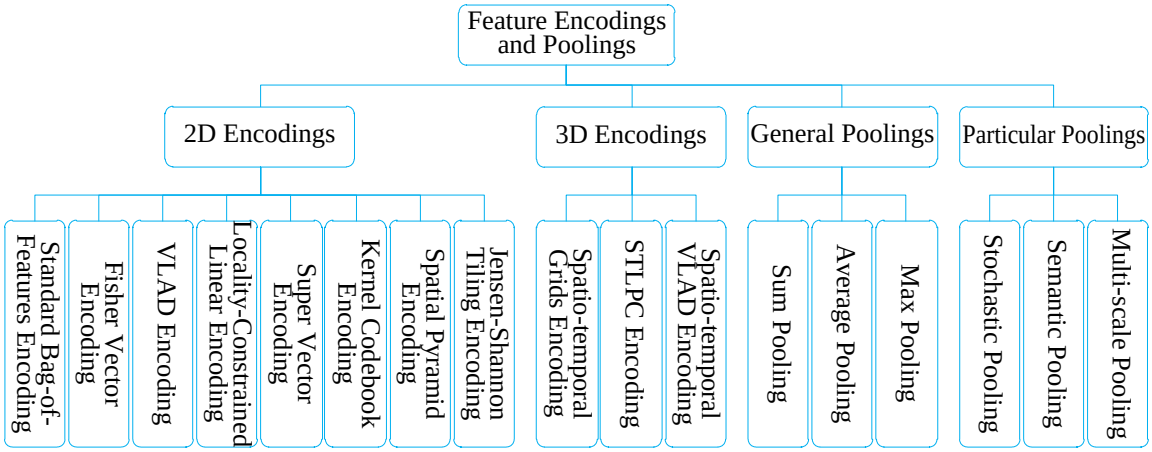

Figure 1. The hierarchical feature encoding and pooling taxonomy of the paper.

Specifically, for 2D encodings, eight popular approaches were surveyed, i.e., standard bag-of-features encoding [49,50], Fisher vector encoding [51–54], VLAD encoding [55], locality-constrained linear encoding [56], super vector encoding [57], kernel codebook encoding [58], spatial pyramid encoding [59,60], and Jensen-Shannon tiling encoding [61]. For 3D encodings, three popular approaches were surveyed, i.e., spatiotemporal grids encoding [62,63], STLPC encoding [64], and spatiotemporal VLAD encoding [65]. For general poolings, three popular approaches were surveyed, i.e., sum pooling [66–70], average pooling [71–76], and max pooling [77–80]. For particular poolings, another three popular approaches were surveyed, i.e., stochastic pooling [81], semantic pooling [82], and multi-scale pooling [83–86].

The rest of the survey paper is organized as follows. Firstly, popular approaches for 2D encodings are surveyed in Section 3. Then, Section 4 covers popular approaches for 3D encodings. In addition, Section 5 presents popular approaches for both general pooling strategies and particular pooling strategies. Finally, Section 6 concludes the survey paper.

3. 2D Encodings

There are eight popular approaches surveyed for 2D encodings, as shown in Table 1.

Table 1. Popular approaches and corresponding references for 2D encodings.

Approaches	References
Standard bag-of-features encoding	Csurka et al., 2004 [49]; Sivic et al., 2003 [50]
Fisher vector encoding	Perronnin et al., 2010 [51]; Perronnin et al., 2007 [52]; Perronnin et al., 2010 [53]; Sánchez et al., 2013 [54]
VLAD encoding	Jégou et al., 2010 [55]
Locality-constrained linear encoding	Wang et al., 2010 [56]
Super vector encoding	Zhou et al., 2010 [57]
Kernel codebook encoding	van Gemert et al., 2008 [58]
Spatial pyramid encoding	Lazebnik et al., 2006 [59]; Grauman et al., 2005 [60]
Jensen-Shannon tiling encoding	Jiang et al., 2014 [61]

3.1. Standard Bag-of-Features Encoding

Since most classifiers require fixed length feature vectors, the ever-changing number of local features in images or videos, e.g., SIFT [87,88] and HOG3D [89], poses difficulties for further pattern recognition tasks. This issue can be solved by the most popular orderless tiling method called bag-of-keypoints [49] or bag-of-visual-words (BOV) [50], which was inspired by the well-known bag-of-words used in text categorization.

BOV treats an image or video frame as a document, the visual vocabulary which is generated by clustering a large set of local features extracted from patches around detected interest points as the word vocabulary, and the cluster centers as the visual words. BOV is a typically high-dimensional sparse visual word occurrence histogram representation.

BOV has been widely used in event and action recognition. Wang and Schmid improved the dense trajectories approach [90] by taking into account camera motion and proposed an efficient state-of-the-art action recognition approach, called Improved Trajectories, where they observed an improvement due to our motion-stabilized descriptors when encoding extracted features with BOV. Lan et al. [91] adopted spatial BOV as one of three encodings to combine more than 10 kinds of features for the TRECVID 2013 multimedia event detection and multimedia event recounting tasks. Ye et al. [92] developed the Raytheon BBN Technologies (BBN) led VISER system for the TRECVID 2013 multimedia event detection and multimedia event recounting tasks, which is based on the BOV approach built on low-level features extracted from pixel patterns in videos.

Although performing surprisingly well, the BOV-based approaches are unable to understand deeply semantic contents of the videos, such as the hierarchical components, which is common issue in high-level event and action recognition [20]. Another drawback is that the important spatial information is lost in this coarse representation.

3.2. Fisher Vector Encoding

Based on the Fisher Kernel framework [93], the Fisher Vector [51–54] is an extension of the bag-of-visual-words (BOV) and encodes feature vectors of patches within one specific image by a gradient vector derived from log likelihood of a universal generative Gaussian mixture model (GMM) [94–99] which can be regarded as a probabilistic visual vocabulary.

Let $X = \{x_t | 1 \leq t \leq T\}$ be the set of T local low level D dimensional feature vectors extracted from an image, e.g., a set of 128 dimensional SIFT feature vectors of interest patches in the image, and $p(x_t|\Theta) = \sum_{i=1}^{N} w_i g(x_t \left| \mu^{(i)}, \Sigma^{(i)} \right.)$ is the probability density function of the GMM, where $\Theta = \{w, \mu, \Sigma\} = \{w_i, \mu_i, \Sigma_i | 1 \leq i \leq N\}$ are parameters, $w_i \in \mathbb{R}$ are the mixture weights with $\sum_{i=1}^{N} w_i = 1$, $\mu^{(i)} \in \mathbb{R}^D$ are the D dimensional mean vectors, $\Sigma^{(i)} \in \mathbb{S}_{++}^D$ are the $D \times D$ symmetric positive definite covariance matrices which are assumed to be diagonal by the variance vector $\sigma_i^2 = \mathrm{diag}\left(\Sigma^{(i)}\right)$, and $g(x_t \left| \mu^{(i)}, \Sigma^{(i)} \right.)$ are the component Gaussian density functions.

Then, the fisher vector encoding adopted in [53] follows steps of:

(1) GMM parameters Θ are trained on a large number of images using the expectation-maximization (EM)-based maximum likelihood estimate (MLE) of [100–102].

(2) Both $\Phi_{\mu_i}^{norm}(X) = \sum_{t=1}^{T} \gamma_t(i)(x_t - \mu_i) / \left(T\sigma_i \sqrt{w_i}\right)$ and $\Phi_{\sigma_i}^{norm}(X) = \sum_{t=1}^{T} \gamma_t(i) \left[(x_t - \mu_i)^2 - \sigma_i^2 \right] / \left(T\sigma_i^2 \sqrt{2w_i}\right)$ are computed, where $1 \leq i \leq N$ and $\gamma_t(i) = (w_i g(x_t|\cdot)) / \sum_{j=1}^{N} w_j g(x_t \left| \cdot \right.)$.

(3) The final fisher vector $\Phi^{norm}(X)$ of the image X is the concatenation of $\Phi_{\mu_i}^{norm}(X)$ and $\Phi_{\sigma_i}^{norm}(X)$ for $1 \leq i \leq N$.

Thus, the dimension of the fisher vector $\Phi^{norm}(X)$ is $2ND$. Furthermore, the authors stated that, linear classifiers provide almost the same results as computation expensive kernel classifiers by using the fisher vector encoding which yields high-classification accuracy and is efficient for large scale processing.

3.3. VLAD Encoding

For indexation and categorization applications, a high-dimensional BOV representation usually shows better yet robust results. However, the efficiency becomes unbearable when those applications are performed on a large amount of data. Jégou et al. [55] proposed a representation, called VLAD (vector of locally aggregated descriptors), for image search on a very large scale. This approach could jointly consider the accuracy, the efficiency, and the computational complexity of the image search.

Let $X = \left\{ x_i \middle| 1 \leq i \leq n, x_i \in \mathbb{R}^d \right\}$ is the set of n local low level d feature vectors extracted from n interest patches of an image, $D = \{w_1, w_2, \cdots, w_k\} \subset \mathbb{R}^{d \times k}$ be the learned codebook with k-means, and $w^* = \underset{w_j \in D, 1 \leq j \leq k}{\operatorname{argmin}} \|x_i - w_j\|$ is the nearest visual word for local patch x_i. Then, the VLAD encoding of the image X can be denoted by a $d \times k$ dimensional vector $f_{VLAD}(X) = (f_{VLAD}(X|w_1), f_{VLAD}(X|w_2), \cdots, f_{VLAD}(X|w_k))$, where $f_{VLAD}(X|w_j) = \sum_{x_i = \underset{1 \leq i \leq n}{\operatorname{argmin}} \|x_i - w_j\|} x_i - w_j$ is the accumulation of differences between the visual word w_j and its nearest multiple image patches, and $1 \leq j \leq k$.

Jégou et al. [55] combined the normalized $f_{VLAD}(X)$ with dimension reduction oriented principal component analysis and indexing oriented approximate nearest neighbor search [103], resulting an accurate, efficient, yet memory friendly image search.

3.4. Locality-Constrained Linear Encoding

Wang et al. [56] pointed out that in order to achieve desirable performance in classification tasks, the most common combinational strategies adopt either linear encodings plus nonlinear classifiers, such as BOV [50] plus χ^2-SVM, or nonlinear encodings plus linear classifiers, such as ScSPM [83] plus linear SVM. Usually, either the nonlinear encodings or the nonlinear classifiers are computationally expensive. As a result, Wang et al. proposed a fast linear encoding approach, called Locality-constrained Linear Coding (LLC) [56] for real-time classification applications.

Suppose $X = \left\{ x_i \middle| 1 \leq i \leq n, x_i \in \mathbb{R}^l \right\}$ is a set of n local l dimensional feature vectors extracted from an image, and $D = \{d_1, d_2, \cdots, d_m\} \subset \mathbb{R}^{l \times m}$ is the encoding dictionary with m words. Then, the process of the proposed fast linear LLC encoding works as follows: (1) For each x_i, find its k nearest neighbors from the dictionary D by using k-NN method, denoted as $D_i = \left\{ v_j \middle| 1 \leq j \leq m \right\}$, where $v_j = d_j$ if v_j is a neighbor of x_i, else $v_j = \mathbf{0}$; (2) Then, for each x_i, its LLC encoding can be obtained from the objective function $f_{LLC}(x_i) = \underset{c_i}{\operatorname{argmin}} \|x_i - c_i^T D_i\|_2^2$, where $\|c_i\|_0 = k$ and $\mathbf{1}^T c_i = 1$.

The authors showed that the promising LLC approach cannot only improve encoding velocity, but also classification performance even with linear classifier than several compared approaches.

3.5. Super Vector Encoding

Zhou et al. [57] proposed a nonlinear high-dimensional sparse coding approach, called Super Vector (SV) encoding, for image representation and classification. The approach is an extended version of the well-known Vector Quantization (VQ) encoding [104].

Let $x \in \mathbb{R}^d$ be the original d dimensional feature vector of an interest patch to be encoded, and $D \subset \mathbb{R}^{d \times l}$ be the encoding dictionary with l words. Then, the SV vector encoded from x can be denoted by $SV_x = \left[\alpha \delta(x|w), \delta(x|w)(x - w)^T \right]^T_{w \in D}$, where α is a nonnegative constant, $w \in \mathbb{R}^l$

represents a word, $\delta(x|w) = 1$ if $w = \underset{w \in D}{\operatorname{argmin}} \|x - w\|$, else $\delta(x|w) = 0$. Thus, the final sparse SV vector is $l(d+1)$ dimensional.

The authors stated that the SV encoding could achieve a lower function approximation error than the VQ encoding, and their proposed classification method achieved state-of-the-art accuracy on the PASCAL dataset.

3.6. Kernel Codebook Encoding

In addition to the loss of spatial structure information, there are another two major drawbacks [58], i.e., codeword uncertainty and codeword plausibility, to the traditional BOV encoding model due to the hard assignment of visual patches to a single codeword. Van Gemert et al. [58] proposed a kernel density estimation approach, called Kernel Codebook encoding, to solve above two issues for scene categorization.

Suppose $X = \left\{ x_j \middle| 1 \le j \le n, x_j \in \mathbb{R}^d \right\}$ is the set of n local low level d dimensional feature vectors extracted from n interest patches of an image, and $D = \{v_1, v_2, \cdots, v_l\} \subset \mathbb{R}^{d \times l}$ be the encoding dictionary with l words. Then, the Kernel codebook encoding of the image X can be denoted by a vector $f_{kcb}(X) = \{f_{dis}(v_i) | 1 \le i \le l\}$, where $f_{dis}(v_i) = \sum_{j=1}^{n} K_\sigma(v_i, x_j)/n$ is the distribution of codeword v_i in the image X, and $K_\sigma(v_i, x_j) = \exp\left\{ -(v_i - x_j)^2 / 2\sigma^2 \right\} / \sqrt{2\pi}\sigma$ is the Gaussian kernel with a scale parameter σ.

Meanwhile, van Gemert et al. [58] concluded that: (1) The encoding $f_{kcb}(X)$ has integrated both codeword uncertainty and codeword plausibility; (2) The Kernel codebook encoding could improve the categorization performance, with either a high-dimensional image descriptor or a smaller dictionary.

3.7. Spatial Pyramid Encoding

Disregarding spatial layouts of interest patches, the bag-of-features-based methods have severely limited descriptive ability in capturing shape information. Lazebnik [59,60] proposed a spatial pyramid encoding and matching approach for scene recognition.

The approach involves repeatedly partitioning the scene image or frame into increasingly fine cells, computing histograms of local features at each cell and measuring similarity with pyramid match kernel. Specifically:

(1) A 400-dimension visual vocabulary is formed by performing k-means clustering on random subset of interest patches, such as 16×16 sized SIFT patches, from the training dataset.
(2) A 4-level spatial pyramid is applied to the scene image or frame X, such that there are 2^l cells along each dimension of level l, resulting in a total of 85 cells in the pyramid, where $0 \le l \le 3$.
(3) For each cell, a 400-dimension bag-of-features histogram is computed. The histograms from all 85 cells are then concatenated and normalized to generate a 34,000-dimension feature vector SP_X as the spatial pyramid encoding of X.
(4) The pyramid match kernel $K(SP_X, SP_Y)$ defined in [59,60] is applied to measure the similarity between scene images or frames X and Y for scene recognition.

Researchers also found that recognition performance from step 1–3 plus SVM classifier with standard kernels is similar to that from step 1–3 plus the pyramid match kernel. The spatial pyramid encoding has been proved effective and is now widely adopted in many applications, such as scene recognition.

3.8. Jensen-Shannon Tiling Encoding

Reseach has shown that representations that consider dynamic salient spatial layouts could always perform better than those with predefined spatial layouts on many recognition or detection tasks [105–107]. However, both the layout of the spatial Bag-of-Words [49,50] encoding and the

layout of spatial pyramid [59,60] encoding are predefined and data independent, which would lead to suboptimal representations.

Jiang et al. [61] proposed a Jensen-Shannon (JS) tiling approach, based on efficiently learned layouts, to encode feature vectors derived from the target image or frame by concatenated histograms from according tiles. Two steps are involved in the JS tiling approach:

(1) It generates systematically all possible tilings, i.e., all vectors $\kappa = \left\langle \kappa_1, \kappa_2, \cdots, \kappa_{|S|} \right\rangle$, from a base mask $S = \left(t_1, t_2, \cdots, t_{|S|}\right)$ by the proposed Algorithm 1, such that the resulting masks can be denoted by $S' = \mathcal{T}_\kappa(S)$, where $\kappa_i \in \mathbb{Z}$ and $\mathcal{T}_\kappa$ is the tiling operator.

(2) It selects the best tiling $\mathcal{T}_\kappa^*$ as the encoding layout, such that $\mathcal{T}_\kappa^* = \underset{\mathcal{T}_\kappa}{\operatorname{argmin}} \cos t(\mathcal{T}_\kappa)$, based on the assumption that an optimal tiling tends to separate positive and negative samples with maximum distance, i.e., JS divergence, where the cost is got by $cost(\mathcal{T}_\kappa) = \lambda|\mathcal{T}_\kappa(S)| - \sum_{j=0}^{|\mathcal{T}_\kappa(S)|-1} JS\left(D_+^j \| D_-^j\right)/|\mathcal{T}_\kappa(S)|$, $|\mathcal{T}_\kappa(S)|$ indicates tile number, λ is a regularized parameter, $JS(\cdot)$ is the JS divergence, $D_+^j = H_+^j/\|H_+^j\|_1$ (or $D_-^j = H_-^j/\|H_-^j\|_1$) is the normalized H_+^j (or H_-^j), $H_+^j = \sum_{y_i=+1}\left(x_i^j + 1\right)/n^+$ (or $H_-^j = \sum_{y_i=-1}\left(x_i^j + 1\right)/n^-$) is average histogram of all positive (or negative) samples for the j^{th} tile, and n^+(or n^-) denotes the total number of positive (or negative) samples.

The authors demonstrated that the JS tiling, as a much faster method, is especially appropriate for large-scale datasets, but with comparable or even better classification results.

4. 3D Encodings

There are three popular approaches surveyed for 3D encodings, as shown in Table 2.

Table 2. Popular approaches and corresponding references for 3D encodings.

Approaches	References
Spatiotemporal grids encoding	Laptev et al., 2008 [62]; Laptev et al., 2007 [63]
STLPC encoding	Shao et al., 2014 [64]
Spatiotemporal VLAD encoding	Duta et al., 2017 [65]

4.1. Spatiotemporal Grids Encoding

Although spatial 2D encodings, such as spatial pyramid and JS tiling, could improve recognition accuracy by considering spatial layouts of interest patches, they are designed mainly for describing one image or one frame and thus not sufficient for describing spatiotemporal targets, such as actions and events.

Laptev et al. [62,63] proposed a spatiotemporal grids (or spatiotemporal bag-of-features) encoding approach for action recognition, which is an extension of the spatial pyramid encoding to spatiotemporal domain. The pipeline involves:

(1) A K-dimension, e.g., $K = 4000$, visual vocabulary is constructed by clustering interest patches sampled from the training videos, such as HOG patches or HOF patches around detected interest points, with k-means algorithm.

(2) For a given test video clip V, 3D interest points are obtained firstly by the spatiotemporal detector, such as Harris 3D detector. Then, spatiotemporal features, such as HOG or HOF, are extracted from patches around those interest points.

(3) The whole test video clip V is divided by a $\sigma_1 \times \sigma_2 \times \tau$ sized grid into cuboids, e.g., spatial sizes $\sigma_1 = 3$, $\sigma_2 = 1$, and temporal size $\tau = 2$. For each cuboid, a K-dimension bag-of-features histogram is formed.

(4) All histograms are further concatenated and normalized to form a $K \times \sigma_1 \times \sigma_2 \times \tau$ dimensional feature vector to represent the video clip V.

The authors concluded that the spatiotemporal grids give a significant gain over the standard bag-of-features methods for action recognition. Besides, the same spatiotemporal grids-based pipeline has been widely used and shown promising results by several research groups [108–111] for multimedia event detection in TRECVID dataset.

4.2. STLPC Encoding

Shao et al. mentioned in paper [64] that local interest points-based sparse representations are not able to preserve adequate spatiotemporal action structures, and traditional tracking or spatial/temporal alignment-based holistic representations are sensitive to background variations. To overcome these defects, they designed a spatiotemporal laplacian pyramid encoding (STLPC) for action recognition. Being different from representation first and encoding or pooling later feature extraction pipelines, Shao et al. proposed to extract action features by successively using frame differences-based STLPC encoding, 3D Gabor filtering representation, and max pooling, where the STLPC encoding involves there primary operations, as follows.

(1) Frame difference volume buliding. For the original video sequence volume V_O, the difference approach is applied as a preprocessing step to generate its frame difference volume V_D.

(2) Spatiotemporal gaussian pyramid buliding. Firstly, generate a four-level frame difference volume pyramid P_{VD} by subsampling the volume V_D with $(1, 1/2, 1/4, 1/8)$ resolution. Then, generate a four-level spatiotemporal gaussian pyramid P_g by convolving a 3-D Gaussian function $f_\sigma(x, y, t) = exp\left(-\left(x^2 + y^2 + t^2\right)/2\sigma^2\right) / \left(\sqrt{2\pi}\sigma\right)^3$ with each level of the pyramid P_{VD}.

(3) Spatiotemporal laplacian pyramid buliding. In the beginning, generate a four-level revised pyramid P_G by expanding each level of the P_g into the same size as of the bottom level. After that, generate a three-level spatiotemporal laplacian pyramid P_L by differencing all two consecutive levels on the revised P_G with $P_L^i = P_G^i - P_G^{i+1}$, where $1 \leq i \leq 3$ is the level number.

Evaluation experiments [64] on four typical action datasets illustrated that the spatiotemporal STLPC, which performed well even with coarse bounding boxes, is an effective yet efficient global encoding for complex human action recognition.

4.3. Spatiotemporal VLAD Encoding

It is well known that spatiotemporal feature extraction is crucial for action recognition. However, as another key factor in action recognition, spatiotemporal encoding has not been paid enough attention. In order to combine both spatial information and temporal information in traditional 2D VLAD encoding, Duta et al. [65] proposed a spatiotemporal VLAD encoding for human action recognition in videos. The pipeline of the ST-VLAD encoding is as follows.

(1) Spatiotemporal deep video features extraction using two stream ConvNet or Improved Dense Trajectories (iDT). For two stream ConvNet, frames are resized to size of 224×224. Then, VGG19 ConvNet is pretrained on ImageNet and then adopted to extract 49 feature vectors with 512 dimentions by each vector for each frame in spatial stream, and VGG16 ConvNet is pretrained on UCF101 and then adopted to extract 49 feature vectors with 512 dimentions feature for each ten frames in temporal stream. For iDT, the resulted dimensionality is 96 for HOG, MBHx and MBHy, and 108 for HOF. Finally, all extracted features are reduced by PCA.

(2) VLAD encoding. Suppose $X = \left\{x_i \middle| 1 \leq i \leq n, x_i \in \mathbb{R}^d\right\}$ is a set of n local low level d dimensional spatial temporal deep video features extracted with approaches in the first step, and $D_a = \{va_1, va_2, \cdots, va_l\} \subset \mathbb{R}^{d \times l}$ is the appearance encoding dictionary with l words. Then, VLAD encoding is defined as a $d \times l$ dimensional feature vector $f_{VLAD}(X) =$

$(f_{VLAD}(X|va_1), \cdots, f_{VLAD}(X|va_l))$, where $f_{VLAD}(X|va_j) = \sum_{x_i \in \mathrm{NN}(va_j)} (x_i - va_j)/|\mathrm{NN}(va_j)|$, and $\mathrm{NN}(va_j) = \underset{1 \le i \le n}{\operatorname{argmin}} \|x_i - va_j\|$.

(3) Spatiotemporal encoding. Suppose $\mathrm{pos}(\cdot)$ is a function to generate a three dimensional normalized position vector, and $D_p = \{vp_1, vp_2, \cdots, vp_m\} \subset \mathbb{R}^{3\times m}$ is the spatiotemporal encoding dictionary with m three dimensional words. Then, the ST encoding is defined as a $m \times (d+l)$ dimensional feature vector by $f_{ST}(X) = (f_{ST}(X|vp_1), \cdots, f_{ST}(X|vp_m))$, where $f_{ST}(X|vp_j) = \sum_{\mathrm{pos}(x_i)\ \&\ \mathrm{pos}(va_k) \in \mathrm{NN}(vp_j)} ((x_i - va_k)/|\mathrm{NN}(va_k)|, MS_i)$, $\mathrm{NN}(vp_j) = \underset{y=\mathrm{pos}(x_i)\ \text{or}\ \mathrm{pos}(va_k)}{\operatorname{argmin}} \|y - vp_j\|$, and MS_i is its membership vector.

(4) ST-VLAD encoding. The final representation, i.e., ST-VLAD, of a video concatenates VLAD encoding and ST encoding in a $d \times l + m \times (d+l)$ dimensional vector.

The authors verified that combing powerful deep features, the proposed ST-VLAD encoding could obtain state-of-art performance on three major challenging action recognition datasets.

5. Pooling Strategies

In computer vision and image processing, there are mainly two applications for pooling strategies. One is to pool the encoding features, and the other is to pool the convolutional features. In this section, let I be an input image. For encoding pooling, suppose $X = \{x_i | 1 \le i \le n, x_i \in \mathbb{R}^r\}$ is the set of n local low level r dimensional feature vectors extracted from I, $E = \{e_j | 1 \le j \le m, e_j \in \mathbb{R}^s\}$ is the encoding feature vectors of I, and $PE = \{pe_k | 1 \le k \le s, pe_k \in \mathbb{R}\}$ be the pooling feature vector of I.

For convolutional pooling, suppose $C = \{c_{i,j} | 1 \le i \le row_c, 1 \le j \le col_c, c_{i,j} \in \mathbb{R}\}$ is the convolutional feature map of I or its upper pooling layer, and $PC = \{pc_{a,b} | 1 \le a \le row_p, 1 \le b \le col_p, pc_{a,b} \in \mathbb{R}\}$ is the pooling feature map of C.

5.1. General Poolings

There are three popular approaches surveyed for general poolings, as shown in Table 3.

Table 3. Popular approaches and corresponding references for general poolings.

Approaches	References
Sum pooling	Peng et al., 2016 [66]; Zhang et al., 2015 [67]; Gao et al., 2016 [68]; LeCun et al., 1998 [69]; Mohedano et al., 2016 [70]
Average pooling	Pinto et al., 2008 [71]; Boureau et al., 2010 [72]; Boureau et al., 2010 [73]; He et al., 2016 [74]; Sainath et al., 2013 [75]; Yu et al., 2014 [76]
Max pooling	Serre et al., 2005 [77]; Sainath et al., 2013 [78]; Scherer et al., 2010 [79]; Wei et al., 2016 [80]

5.1.1. Sum Pooling

For encoding pooling, mathematical representation of the sum pooling can be expressed by: $pe_k = \sum_{j=1}^{m} e_{j,k}$. Although sum pooling is an intuitive pooling strategy, there are a number of papers using this method. For example, Peng et al. [66] analyzed action recognition performance among several bag of visual words and fusion methods, where they adopted sum pooling and power l_2-normalization for pooling and normalization strategy. Zhang et al. [67] gave a probabilistic interpretation why the max pooling was usually better than sum pooling in the context of sparse coding framework for image retrieval applications, since max pooling tended to increase the discrimination of the similarity measurement than sum pooling. Besides, they proposed a modified sum pooling method, improving the retrieval accuracy significantly over the max pooling strategy.

For convolutional pooling, sum pooling can be derived by: $pc_{a,b} = \sum_{(i,j)\in\mathcal{R}_{a,b}} c_{i,j}$. There are some attempts using sum pooling for convolutional neural network (CNN) based applications. For instance, Gao et al. [68] proposed a compact bilinear pooling method for image classification based on a kernelized analysis of bilinear sum pooling. They verified that the method could reduce the feature dimensionality two orders of magnitude with little loss in performance, and the CNN back-propagation can be efficiently computed. LeCun et al. [69] developed a typical CNN, called LeNet-5, for isolated character recognition. The LeNet-5 architecture consisted of two convolution layers, two sum-pooling layers, and several full connection layers. However, Mohedano et al. [70] found that for instance retrieval tasks, even bag-of-words aggregation could outperform techniques using sum pooling when combing with local CNN features at the challenging TRECVID INS benchmark.

5.1.2. Average Pooling

For encoding pooling, the average pooling strategy can be represented by: $pe_k = \sum_{j=1}^{m} e_{j,k}/m$. There are many works using average pooling. For instance, Pinto et al. [71] constructed a biologically inspired object recognition system, i.e., a simple V1-like model, based on average pooling [72]. This model outperforms state-of-the-art object recognition systems on a standard natural image recognition test. However, many researchers have shown that, in most encoding pooling-based vision applications, average pooling usually was not the best pooling strategy. For example, Boureau et al. [73] provided a theoretical and empirical insight into the performance between max pooling and average pooling. They pointed out that max pooling outperformed almost always average pooling, especially dramatically when using a linear SVM.

For convolutional pooling, average pooling can be denoted as: $pc_{a,b} = \sum_{(i,j)\in\mathcal{R}_{a,b}} c_{i,j}/\left|\mathcal{R}_{a,b}\right|^2$, where $\mathcal{R}_{a,b}$ is a pooling region, and $\left|\mathcal{R}_{a,b}\right|$ is its edge length. He et al. [74] introduced a deep residual net, called ResNet, for large-scale image recognition. The ResNet ended with a global average pooling layer and a fully connected layer with softmax. The authors won first place in several tracks in ILSVRC & COCO 2015 competitions by using the ResNet. For convolutional pooling-based vision applications, the situation that average pooling is always not the best choice is similar. For example, Sainath et al. [75] explored four pooling strategies in frequency only for an LVCSR speech task, and concluded that either l_p pooling or Stochastic pooling can address the issues of max pooling or average pooling. Yu et al. [76] invented a mixed pooling method to regularize CNN. They demonstrated that the mixed pooling method is superior to both max pooling and average pooling, since the latter may reduce largely the feature map if there are many zero elements. In 2016 TRECVID competition, the CMU Informedia team [112] adopted the average-pooling for video representation in the multimedia event detection task, the Ad-hoc video search task, and the surveillance event detection task. Besides, they also adopted the average-pooling for document representation in the video hyperlinking task. Their experiments indicated that the average-pooling is an ideal discriminative strategy for hybrid representations.

5.1.3. Max Pooling

As one of the most welcome pooling strategies in vision related tasks, max pooling has been widely used in both encoding-based and convolutional-based pooling applications. For encoding pooling, the max pooling strategy has following formalization: $pe_k = \max_{j\in\{1,2\cdots,m\}} e_{j,k}$. This max-pooling has been empirically justified by many algorithms. For example, Serre et al. [77] developed a biologically motivated framework. The framework adopts only two major kinds of computations, i.e., template matching and max pooling, to obtain a set of scale- and translation-invariant C2 features, for robust object recognition. Boureau et al. [72] conducted a theoretical analysis of average pooling and max pooling for visual recognition tasks. The authors concluded that the recognition performance using

pooling strategies can be influenced by many factors, such as sample cardinality, resolution, codebook size, and max pooling performs no worse than average pooling in most cases.

For convolutional pooling, the max pooling strategy can be formalized as: $pc_{a,b} = \max_{(i,j)\in\mathcal{R}_{a,b}} c_{i,j}$, where $\mathcal{R}_{a,b}$ is a pooling region. There are also lots of researches using max pooling in CNN to generate deep features. For example, Sainath et al. [78] explored applying CNNs to large-vocabulary speech tasks, and showed that their convolutional network architecture, which consisted of a convolutional and max-pooling layer, was an improved CNN. Scherer et al. [79] evaluated two pooling operations in convolutional architectures for object recognition, and showed that a maximum pooling operation significantly outperformed a subsampling operation. Wei et al. [80] presented a flexible CNN framework, which can be pre-trained well on large-scale single-label image datasets, for multi-label image classification. The framework generated its ultimate multi-label predictions with a cross-hypothesis max-pooling operation on confidence vectors obtained from the input hypotheses using the shared CNN.

5.2. Particular Poolings

There are three popular approaches surveyed for particular poolings, as shown in Table 4.

Table 4. Popular approaches and corresponding references for particular poolings.

Approaches	References
Stochastic pooling	Zeiler et al., 2013 [81]
Semantic pooling	Chang et al., 2017 [82]
Multi-scale pooling	Yang et al., 2009 [83]; Gong et al., 2014 [84]; He et al., 2015 [85]; Szegedy et al., 2015 [86]

5.2.1. Stochastic Pooling

For large-scale convolutional neural networks, how to simultaneously reduce computational complexity and keep visual invariance in training processes has become an important issue. Conventionally, researchers solved this issue by adding several extra average pooling or max pooling layers. However, both types of pooling have drawbacks. Namely, average pooling leads to small-pooled responses, and max pooling leads to over-fitting. Thus, Zeiler et al. [81] proposed a stochastic pooling approach for regularization of deep convolutional neural networks.

Firstly, they derive a probability map $PM = \{pm_{i,j}\}$ from the convolutional feature map C using $pm_{i,j} = c_{i,j} / \sum_{(p,q)\in\mathcal{R}_{i,j}} c_{p,q}$, where $\mathcal{R}_{i,j}$ is the pooling region determined by $c_{i,j}$. Secondly, they compute the pooling feature map $PC = \{pc_{a,b}\}$ using $pc_{a,b} = c_{i^*,j^*}$, $s.t.$ $(i^*, j^*) \sim P_{MN}\left(pm_{1,1}, \cdots, pm_{|\mathcal{R}_{i,j}|_h,|\mathcal{R}_{i,j}|_v}\right)$, where (i^*, j^*) is pooled activation, $P_{MN}(\cdot)$ is the multinomial distribution, and $\left|\mathcal{R}_{i,j}\right|_h$ or $\left|\mathcal{R}_{i,j}\right|_v$ refers to horizontal or vertical dimension of the pooling region.

Thus, the convolutional feature map in the large CNN can be randomly pooled according to the multinomial distribution. The authors also stated that the simple yet effective stochastic pooling strategy can be combined with any other forms of regularization to prevent over-fitting and reduce computational complexity for deep convolutional neural networks based applications.

5.2.2. Semantic Pooling

For complex event detection in long internet videos with few relevant shots, traditional pooling strategies treat usually each shot equally and cannot aggregate the shots based on their relevance with respect to the event of interest [82]. Chang et al. [82] proposed a semantic pooling approach to prioritize CNN shot outputs according to their semantic saliencies.

Firstly, shots-based CNN feature extraction. Specifically, compute the average number of key frames m for all videos in the experiment dataset, and adopt the color histogram difference-based

shot boundary algorithm to divide each video into m shots, denoted by $V = \{SH_i | 1 \leq i \leq m\}$. Then, select randomly one frame in the shot as its key frame, and extract CNN features on all key frames, denoted by $XC = \left\{ xc_i \middle| 1 \leq i \leq m, xc_i \in \mathbb{R}^d \right\}$.

Secondly, concept probability-based feature extraction. Specifically, apply common datasets of action and event recognition to train beforehand plenty of semantic auxiliary concept detectors, i.e., $c = 1534$, and generate a c dimensional probability vector for each shot by concatenating responses of all concept detectors on the shot, denoted by $XV = \{xv_i | 1 \leq i \leq m, xv_i \in \mathbb{R}^c\}$.

Thirdly, concept relevance-based feature extraction. Specifically, bring the English Wikipedia dump to train previously a skip-gram model, and employ the skip-gram model together with the Fisher vector encoding approach to vectorize both the textual event descriptions and the c concept names. Then, compute the cosine distance between the overall textual description vector and each concept name vector, resulting in a concept relevance vector $XR = \{xr_i | 1 \leq i \leq c, xr_i \in \mathbb{R}\}$.

Subsequently, saliency-based semantic pooling. For each shot i, compute a semantic saliency score sa_i inner product between its corresponding row vector xv_i of the concept probability matrix XV and the concept relevance vector XR, producing a semantic saliency score vector $SA = \left\{ sa_i \middle| sa_i = XR^T xv_i, 1 \leq i \leq m, sa_i \in \mathbb{R} \right\}$. Then, rank all CNN feature vector xc_i in a descending order according to its saliency score sa_i, and concatenate successively the ranked feature vectors as the final deep representation.

In order to exploit ordering information in the semantic pooling-based deep features, the authors [82] designed also a nearly isotonic classification approach, and verified through a number of experiments that the combination of the flexible deep representation and the sophisticated classification exhibited higher discriminative power in event analysis tasks of event detection, event recognition, and event recounting.

5.2.3. Multi-Scale Pooling

For encoding pooling, Yang et al. [83] proposed a multi-scale spatial max pooling approach to generate nonlinear features based on sparse coding, as a generalized vector quantization, for fast yet accurate image classification.

For convolutional pooling, Gong et al. [84] proposed a multi-scale orderless pooling approach to improve geometric invariance in CNN for classification and matching of highly variable scenes. He et al. [85] proposed a spatial pyramid pooling approach to generate fixed length representations for CNN-based visual recognition. The major contribution of the spatial pyramid pooling approach is its additional three-scale max-pooling layer. Szegedy et al. [86] proposed a 22-layer deep model, namely GoogLeNet, for classification and detection in the ImageNet challenge competition. The GoogLeNet employs a parallel multi-scale hybrid pooling architecture to reduce computing resources in deeper convolutions.

6. Conclusions

In this paper, we have surveyed comprehensively the approaches of encodings and poolings that have been previously studied for feature representations in action and event recognition on uncontrolled video clips, and summarized systematically both underlying theoretical principles and original experimental conclusions of those approaches. Furthermore, we have designed an approach-based taxonomy to categorize the most popular previous research work on encodings and poolings by 2D encodings, 3D encodings, general poolings, and particular poolings. As mentioned above, feature encoding and feature pooling are only two of three key components in feature representation approach, whereas the feature representation approach is only one of three key components in action and event recognition. In the future, we will conduct three more surveys for the other three key components, namely a survey on feature extraction approaches, a survey on pattern recognition models, and a survey on performance evaluation strategies, for recognition of complex actions and events.

Acknowledgments: This paper was supported by the Natural Science Foundation of Guangdong Province, China under Grant No. 2017A030313373, the Science and Technology Planning Project of Guangdong Province, China under Grant Nos. 2016A020210103 and 2017A020208054, Social Science Planning Project of Guangzhou, China under Grant No. 2017GZQN05, the Scientific Research Fund of SiChuan Provincial Education Department, China under Grant No. 17ZA0327, the Basic Theory and Scientific Research Project of Jiangmen City under title "Evolution mechanism of some social networks and empirical study", and the Doctor Startup Foundation of Wuyi University under Grant No. 2014BS07. We would like to thank anonymous reviewers for helpful comments.

Author Contributions: Changyu Liu designed the hierarchical taxonomy for state of the art approaches in action and event detection, and wrote the survey manuscript; Bin Lu and Qian Zhang performed the technical revision; Cong Li conducted the grammar revision.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Coşar, S.; Donatiello, G.; Bogorny, V.; Garate, C.; Alvares, L.O.; Brémond, F. Toward abnormal trajectory and event detection in video surveillance. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 683–695. [CrossRef]
2. García-Martín, Á.; Martínez, J.M. People detection in surveillance: Classification and evaluation. *IET Comput. Vis.* **2015**, *9*, 779–788. [CrossRef]
3. Lee, S.C.; Nevatia, R. Hierarchical abnormal event detection by real time and semi-real time multi-tasking video surveillance system. *Mach. Vis. Appl.* **2014**, *25*, 133–143. [CrossRef]
4. Fang, Z.; Fei, F.; Fang, Y.; Lee, C.; Xiong, N.; Shu, L.; Chen, S. Abnormal event detection in crowded scenes based on deep learning. *Multimedia Tools Appl.* **2016**, *75*, 14617–14639. [CrossRef]
5. Chang, X.; Ma, Z.; Lin, M.; Yang, Y.; Hauptmann, A. Feature interaction augmented sparse learning for fast Kinect motion detection. *IEEE Trans. Image Process.* **2017**, *26*, 3911–3920. [CrossRef] [PubMed]
6. Morariu, V.I.; Davis, L.S. Multi-agent event recognition in structured scenarios. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 3289–3296.
7. Rautaray, S.S.; Agrawal, A. Interaction with virtual game through hand gesture recognition. In Proceedings of the 2011 International Conference on Multimedia, Signal Processing and Communication Technologies, Aligarh, India, 17–19 December 2011; pp. 244–247.
8. Fothergill, S.; Mentis, H.; Kohli, P.; Nowozin, S. Instructing people for training gestural interactive systems. In Proceedings of the 30th ACM Conference on Human Factors in Computing Systems, Austin, TX, USA, 5–10 May 2012; pp. 1737–1746.
9. Lin, F.; Wang, A.; Cavuoto, L.; Xu, W. Toward unobtrusive patient handling activity recognition for injury reduction among at-risk caregivers. *IEEE J. Biomed. Health Inform.* **2017**, *21*, 682–695. [CrossRef] [PubMed]
10. Tripathy, A.K.; Carvalho, R.; Pawaskar, K.; Yadav, S.; Yadav, V. Mobile based healthcare management using artificial intelligence. In Proceedings of the International Conference on Technologies for Sustainable Development, Mumbai, India, 4–6 February 2015; pp. 1–6.
11. Jalal, A.; Uddin, M.Z.; Kim, J.T.; Kim, T.S. Daily human activity recognition using depth silhouettes and ℜ transformation for smart home. In Proceedings of the 9th International Conference on Smart Homes and Health Telematics: Toward Useful Services for Elderly and People with Disabilities, Montreal, QC, Canada, 20–22 June 2011; pp. 25–32.
12. Bradbury-Jones, C.; Taylor, J.; Kroll, T.; Duncan, F. Domestic abuse awareness and recognition among primary healthcare professionals and abused women: A qualitative investigation. *J. Clin. Nurs.* **2014**, *23*, 3057–3068. [CrossRef] [PubMed]
13. Niu, L.; Xu, X.; Chen, L.; Duan, L.; Xu, D. Action and event recognition in videos by learning from heterogeneous web sources. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *28*, 1290–1304. [CrossRef] [PubMed]
14. Yu, L.; Yang, Y.; Huang, Z.; Wang, P.; Song, J.; Shen, H.T. Web video event recognition by semantic analysis from ubiquitous documents. *IEEE Trans. Image Process.* **2016**, *25*, 5689–5701. [CrossRef] [PubMed]
15. Jalal, A.; Zeb, M.A. Security enhancement for e-learning portal. *Int. J. Comput. Sci. Netw. Secur.* **2008**, *8*, 41–45.
16. Ladan, M.I. E-Commerce security issues. In Proceedings of the 2014 International Conference on Future Internet of Things and Cloud, Barcelona, Spain, 27–29 August 2014; pp. 197–201.

17. Jalal, A.; Shahzad, A. Multiple facial feature detection using vertex-modeling structure. In Proceedings of the International Conference on Interactive Computer Aided Learning, Villach, Austria, 26–28 September 2007; pp. 1–7.
18. Jalal, A. Security architecture for third generation (3G) using GMHS cellular network. In Proceedings of the 3rd International Conference on Emerging Technologies, Islamabad, Pakistan, 12–13 November 2007; pp. 74–79.
19. Over, P.; Fiscus, J.; Sanders, G.; Joy, D.; Michel, M.; Awad, G.; Smeaton, A.; Kraaij, W.; Quénot, G. TRECVID 2012—An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics. Available online: http://www-nlpir.nist.gov/projects/tvpubs/tv12.papers/tv12overview.pdf (accessed on 26 October 2017).
20. Jiang, Y.G.; Bhattacharya, S.; Chang, S.F.; Shah, M. High-level event recognition in unconstrained videos. *Int. J. Multimedia Inf. Retr.* **2013**, *2*, 73–101. [CrossRef]
21. Lara, O.D.; Labrador, M.A. A survey on human activity recognition using wearable sensors. *IEEE Commun. Surv. Tutor.* **2013**, *15*, 1192–1209. [CrossRef]
22. Jalal, A.; Kim, J.T.; Kim, T.S. Development of a life logging system via depth imaging-based human activity recognition for smart homes. In Proceedings of the 8th International Symposium on Sustainable Healthy Buildings, Seoul, Korea, 19 September 2012; pp. 91–95.
23. Yang, A.Y.; Iyengar, S.; Kuryloski, P.; Jafari, R. Distributed segmentation and classification of human actions using a wearable motion sensor network. In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
24. Song, Y.; Tang, J.; Liu, F.; Yan, S. Body surface context: A new robust feature for action recognition from depth videos. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *24*, 952–964. [CrossRef]
25. Jalal, A.; Sarif, N.; Kim, J.T.; Kim, T.S. Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home. *Indoor Built Environ.* **2013**, *22*, 271–279. [CrossRef]
26. Althloothi, S.; Mahoor, M.H.; Zhang, X.; Voyles, R.M. Human activity recognition using multi-features and multiple kernel learning. *Pattern Recognit.* **2014**, *47*, 1800–1812. [CrossRef]
27. Jalal, A.; Kamal, S.; Kim, D. Shape and motion features approach for activity tracking and recognition from kinect video camera. In Proceedings of the IEEE 29th International Conference on Advanced Information Networking and Applications Workshops, Gwangju, Korea, 25–27 March 2015; pp. 445–450.
28. Jalal, A.; Kamal, S.; Kim, D. A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. *Sensor* **2014**, *14*, 11735–11759. [CrossRef] [PubMed]
29. Kamal, S.; Azurdia-Meza, C.A.; Lee, K. Family of Nyquist-I pulses to enhance orthogonal frequency division multiplexing system performance. *IETE Tech. Rev.* **2016**, *33*, 187–198. [CrossRef]
30. Li, Y.; Zhang, X.; Jin, H.; Li, X.; Wang, Q.; He, Q.; Huang, Q. Using multi-stream hierarchical deep neural network to extract deep audio feature for acoustic event detection. *Multimedia Tools Appl.* **2017**. [CrossRef]
31. Kumar, A.; Raj, B. Audio event and scene recognition: A unified approach using strongly and weakly labeled data. In Proceedings of the 2017 International Joint Conference on Neural Networks, Anchorage, AK, USA, 14–19 May 2017; pp. 3475–3482.
32. Farooq, A.; Jalal, A.; Kamal, S. Dense RGB-D map-based human tracking and activity recognition using skin joints features and self-organizing map. *KSII Trans. Internet Inf. Syst.* **2015**, *9*, 1856–1869.
33. Siswanto, A.R.S.; Nugroho, A.S.; Galinium, M. Implementation of face recognition algorithm for biometrics based time attendance system. In Proceedings of the 2014 International Conference on ICT for Smart Society, Bandung, Indonesia, 23–24 September 2014; pp. 149–154.
34. Manwatkar, P.M.; Yadav, S.H. Text recognition from images. In Proceedings of the 2015 IEEE International Conference on Innovations in Information, Embedded and Communication Systems, Coimbatore, India, 19–20 March 2015; pp. 1–6.
35. Chang, X.; Ma, Z.; Yang, Y.; Zeng, Z.; Hauptmann, A.G. Bi-level semantic representation analysis for multimedia event detection. *IEEE Trans. Cybern.* **2017**, *47*, 1180–1197. [CrossRef] [PubMed]
36. Jalal, A.; Kamal, S.; Kim, D. Depth Silhouettes Context: A new robust feature for human tracking and activity recognition based on embedded HMMs. In Proceedings of the 2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence, Goyang City, Korea, 28–30 October 2015; pp. 294–299.
37. Kamal, S.; Jalal, A.; Kim, D. Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM. *J. Electr. Eng. Technol.* **2016**, *11*, 1857–1862. [CrossRef]

38. Jalal, A.; Kim, Y.H.; Kim, Y.J.; Kamal, S.; Kim, D. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognit.* **2017**, *61*, 295–308. [CrossRef]
39. Vishwakarma, S.; Agrawal, A. A survey on activity recognition and behavior understanding in video surveillance. *Vis. Comput.* **2013**, *29*, 983–1009. [CrossRef]
40. Guo, G.; Lai, A. A survey on still image based human action recognition. *Pattern Recognit.* **2014**, *47*, 3343–3361. [CrossRef]
41. Aggarwal, J.K.; Xia, L. Human activity recognition from 3D data: A review. *Pattern Recognit. Lett.* **2014**, *48*, 70–80. [CrossRef]
42. Ziaeefard, M.; Bergevin, R. Semantic human activity recognition: A literature review. *Pattern Recognit.* **2015**, *48*, 2329–2345. [CrossRef]
43. Akoglu, L.; Tong, H.; Koutra, D. Graph based anomaly detection and description: A survey. *Data Min. Knowl. Discov.* **2015**, *29*, 626–688. [CrossRef]
44. Zhen, X.; Shao, L. Action recognition via spatio-temporal local features: A comprehensive study. *Image Vis. Comput.* **2016**, *50*, 1–13. [CrossRef]
45. Xu, H.; Tian, Q.; Wang, Z.; Wu, J. A survey on aggregating methods for action recognition with dense trajectories. *Multimedia Tools Appl.* **2016**, *75*, 5701–5717. [CrossRef]
46. Zhu, F.; Shao, L.; Xie, J.; Fang, Y. From handcrafted to learned representations for human action recognition: A survey. *Image Vis. Comput.* **2016**, *55*, 42–52. [CrossRef]
47. Wu, D.; Sharma, N.; Blumenstein, M. Recent advances in video-based human action recognition using deep learning: A review. In Proceedings of the International Joint Conference on Neural Networks, Anchorage, AK, USA, 14–19 May 2017; pp. 2865–2872.
48. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. *Image Vis. Comput.* **2017**, *60*, 4–21. [CrossRef]
49. Csurka, G.; Dance, C.; Fan, L.; Willamowski, J.; Bray, C. Visual categorization with bags of keypoints. In Proceedings of the ECCV International Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic, 11–14 May 2004; pp. 1–16.
50. Sivic, J.; Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; pp. 1470–1477.
51. Perronnin, F.; Liu, Y.; Sánchez, J.; Poirier, H. Large-scale image retrieval with compressed fisher vectors. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3384–3391.
52. Perronnin, F.; Dance, C. Fisher kernels on visual vocabularies for image categorization. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
53. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In Proceedings of the 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 143–156.
54. Sánchez, J.; Perronnin, F.; Mensink, T.; Verbeek, J. Image classification with the Fisher vector: Theory and practice. *Int. J. Comput. Vis.* **2013**, *105*, 222–245. [CrossRef]
55. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
56. Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T.; Gong, Y. Locality-constrained linear coding for image classification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3360–3367.
57. Zhou, X.; Yu, K.; Zhang, T.; Huang, T.S. Image classification using super-vector coding of local image descriptors. In Proceedings of the 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 141–154.
58. Van Gemert, J.C.; Geusebroek, J.M.; Veenman, C.J.; Smeulders, A.W.M. Kernel codebooks for scene categorization. In Proceedings of the 10th European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 696–709.

59. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categorie. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; pp. 2169–2178.
60. Grauman, K.; Darrell, T. The pyramid match kernel: Discriminative classification with sets of image features. In Proceedings of the IEEE International Conference on Computer Vision, Beijing, China, 17–20 October 2005; pp. 1458–1465.
61. Jiang, L.; Tong, W.; Meng, D.; Hauptmann, A.G. Towards efficient learning of optimal spatial bag-of-words representations. In Proceedings of the ACM International Conference on Multimedia Retrieval 2014, Glasgow, UK, 1–4 April 2014; pp. 121–128.
62. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning realistic human actions from movies. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
63. Laptev, I.; Pérez, P. Retrieving actions in movies. In Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
64. Shao, L.; Zhen, X.; Tao, D.; Li, X. Spatio-temporal Laplacian pyramid coding for action recognition. *IEEE Trans. Cybern.* **2014**, *44*, 817–827. [CrossRef] [PubMed]
65. Duta, I.C.; Ionescu, B.; Aizawa, K.; Sebe, N. Spatio-temporal VLAD encoding for human action recognition in videos. In Proceedings of the International Conference on Multimedia Modeling, Reykjavik, Iceland, 4–6 January 2017; pp. 365–378.
66. Peng, X.; Wang, L.; Wang, X.; Qiao, Y. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Comput. Vis. Image Underst.* **2016**, *150*, 109–125. [CrossRef]
67. Zhang, Y.; Chen, J.; Huang, X.; Wang, Y. A probabilistic analysis of sparse coded feature pooling and its application for image retrieval. *PLoS ONE* **2015**, *10*, e0131721. [CrossRef] [PubMed]
68. Gao, Y.; Beijbom, O.; Zhang, N.; Darrell, T. Compact bilinear pooling. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 317–326.
69. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
70. Mohedano, E.; McGuinness, K.; O'Connor, N.E.; Salvador, A.; Marqués, F.; Giró-i-Nieto, X. Bags of local convolutional features for scalable instance search. In Proceedings of the 2016 ACM International Conference on Multimedia Retrieval, New York, NY, USA, 6–9 June 2016; pp. 327–331.
71. Pinto, N.; Cox, D.; DiCarlo, J. Why is real-world visual object recognition hard. *PLoS Comput. Biol.* **2008**, *4*, 151–156. [CrossRef] [PubMed]
72. Boureau, Y.L.; Ponce, J.; LeCun, Y. A theoretical analysis of feature pooling in visual recognition. In Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010; pp. 111–118.
73. Boureau, Y.L.; Bach, F.; LeCun, Y.; Ponce, J. Learning mid-level features for recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2559–2566.
74. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
75. Sainath, T.N.; Kingsbury, B.; Mohamed, A.; Dahl, G.E.; Saon, G.; Soltau, H.; Beran, T.; Aravkin, A.Y.; Ramabhadran, B. Improvements to deep convolutional neural networks for LVCSR. In Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 8–13 December 2013; pp. 315–320.
76. Yu, D.; Wang, H.; Chen, P.; Wei, Z. Mixed pooling for convolutional neural networks. In Proceedings of the 9th International Conference on Rough Sets and Knowledge Technology, Shanghai, China, 24–26 October 2014; pp. 364–375.
77. Serre, T.; Wolf, L.; Poggio, T. Object recognition with features inspired by visual cortex. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 994–1000.

78. Sainath, T.N.; Mohamed, A.; Kingsbury, B.; Ramabhadran, B. Deep convolutional neural networks for LVCSR. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8614–8618.
79. Scherer, D.; Müller, A.; Behnke, S. Evaluation of pooling operations in convolutional architectures for object recognition. In Proceedings of the 20th International Conference on Artificial Neural Networks, Thessaloniki, Greece, 15–18 September 2010; pp. 92–101.
80. Wei, Y.; Xia, W.; Lin, M.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; Yan, S. HCP: A flexible CNN framework for multi-label image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1901–1907. [CrossRef] [PubMed]
81. Zeiler, M.D.; Fergus, R. Stochastic pooling for regularization of deep convolutional neural networks. In Proceedings of the International Conference on Learning Representation, Scottsdale, AZ, USA, 2–4 May 2013; pp. 1–9.
82. Chang, X.J.; Yu, Y.L.; Yang, Y.; Xing, E.P. Semantic pooling for complex event analysis in untrimmed videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1617–1632. [CrossRef] [PubMed]
83. Yang, J.; Yu, K.; Gong, Y.; Huang, T. Linear spatial pyramid matching using sparse coding for image classification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1794–1801.
84. Gong, Y.; Wang, L.; Guo, R.; Lazebnik, S. Multi-scale orderless pooling of deep convolutional activation features. In Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 392–407.
85. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
86. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
87. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; pp. 1150–1157.
88. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
89. Kläser, A.; Marszałek, M.; Schmid, C. A spatio-temporal descriptor based on 3d-gradients. In Proceedings of the British Machine Vision Conference 2008, Leeds, UK, 1–4 September 2008; pp. 1–10.
90. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 3551–3558.
91. Lan, Z.; Jiang, L.; Yu, S.I.; Gao, C.; Rawat, S.; Cai, Y.; Xu, S.; Shen, H.; Li, X.; Wang, Y.; et al. Informedia @ TRECVID 2013. Available online: http://www-nlpir.nist.gov/projects/tvpubs/tv13.papers/informedia.pdf (accessed on 26 October 2017).
92. Luisier, F.; Tickoo, M.; Andrews, W.; Ye, G.; Liu, D.; Chang, S.F.; Salakhutdinov, R.; Morariu, V.; Davis, L.; Gupta, A.; et al. BBN VISER TRECVID 2013 Multimedia Event Detection and Multimedia Event Recounting Systems. Available online: http://www-nlpir.nist.gov/projects/tvpubs/tv13.papers/bbnviser.pdf (accessed on 26 October 2017).
93. Jaakkola, T.; Haussler, D. Exploiting generative models in discriminative classifiers. In Proceedings of the 12th Annual Conference on Neural Information Processing Systems, Denver, CO, USA, 30 November–5 December 1998; pp. 487–493.
94. Chellappa, R.; Veeraraghavan, A.; Ramanathan, N.; Yam, C.Y.; Nixon, M.S.; Elgammal, A.; Boyd, J.E.; Little, J.J.; Lynnerup, N.; Larsen, P.K.; et al. Chapter 533: Gaussian mixture models. In *Encyclopedia of Biometrics*; Li, S.Z., Jain, A., Eds.; Springer: New York, NY, USA, 2009; pp. 659–663.
95. Yu, G.; Sapiro, G.; Mallat, S. Solving inverse problems with piecewise linear estimators: From Gaussian mixture models to structured sparsity. *IEEE Trans. Image Process.* **2012**, *21*, 2481–2499. [PubMed]
96. Jian, B.; Vemuri, B.C. Robust point set registration using gaussian mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1633–1645. [CrossRef] [PubMed]
97. Kerroum, M.A.; Hammouch, A.; Aboutajdine, D. Textural feature selection by joint mutual information based on Gaussian mixture model for multispectral image classification. *Pattern Recognit. Lett.* **2010**, *31*, 1168–1174. [CrossRef]

98. Duda, R.O.; Hart, P.E.; Storck, D.J. *Pattern Classification*, 2nd ed.; John Wiley and Sons, Inc.: Hoboken, NJ, USA, 2003.
99. Lin, H.H.; Chuang, J.H.; Liu, T.L. Regularized background adaptation: A novel learning rate control scheme for Gaussian mixture modeling. *IEEE Trans. Image Process.* **2011**, *20*, 822–836. [PubMed]
100. Perronnin, F.; Dance, C.; Csurka, G.; Bressan, M. Adapted vocabularies for generic visual categorization. In Proceedings of the 9th European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 464–475.
101. McLachlan, G.J.; Basford, K.E. *Mixture Models: Inference and Applications to Clustering (Statistics: Textbooks & Monographs)*; CRC Press: Boca Raton, FL, USA, 1988.
102. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–38.
103. Jegou, H.; Douze, M.; Schmid, C. Product quantization for nearest neighbor search. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 117–128. [CrossRef] [PubMed]
104. Gray, R. Vector quantization. *IEEE ASSP Mag.* **1984**, *1*, 4–29. [CrossRef]
105. Tong, W.; Yang, Y.; Jiang, L.; Yu, S.I.; Lan, Z.; Ma, Z.; Hauptmann, A.G. E-LAMP: Integration of innovative ideas for multimedia event detection. *Mach. Vis. Appl.* **2014**, *25*, 5–15. [CrossRef]
106. Cai, Y.; Chen, Q.; Brown, L.; Datta, A.; Fan, Q.; Feris, R.; Yan, S.; Hauptmann, A.; Pankanti, S. CMU-IBM-NUS @ TRECVID 2012: Surveillance Event Detection (SED). Available online: http://www-nlpir.nist.gov/projects/tvpubs/tv12.slides/tv12.cmu.sed.slides.pdf (accessed on 26 October 2017).
107. Cao, L.; Chang, S.F.; Codella, N.; Cotton, C.; Ellis, D.; Gong, L.; Hill, M.; Hua, G.; Kender, J.; Merler, M.; et al. IBM Research and Columbia University TRECVID-2011 Multimedia Event Detection (MED) System. Available online: http://www-nlpir.nist.gov/projects/tvpubs/tv11.papers/ibm.pdf (accessed on 26 October 2017).
108. Yang, X.; Liu, Z.; Zavesky, E.; Gibbon, D.; Shahraray, B.; Tian, Y. AT&T Research at TRECVID 2013: Surveillance Event Detection. Available online: http://www-nlpir.nist.gov/projects/tvpubs/tv13.papers/att.pdf (accessed on 26 October 2017).
109. Merler, M.; Huang, B.; Xie, L.; Hua, G.; Natsev, A. Semantic model vectors for complex video event recognition. *IEEE Trans. Multimedia* **2012**, *14*, 88–101. [CrossRef]
110. Xian, Y.; Rong, X.J.; Yang, X.D.; Tian, Y.L. CCNY at TRECVID 2014: Surveillance Event Detection. Available online: http://www-nlpir.nist.gov/projects/tvpubs/tv14.papers/ccny.pdf (accessed on 26 October 2017).
111. Shi, F.; Petriu, E.; Laganiere, R. Sampling strategies for real-time action recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2595–2602.
112. Liang, J.W.; Chen, J.; Huang, P.Y.; Li, X.C.; Jiang, L.; Lan, Z.Z.; Pan, P.B.; Fan, H.H.; Jin, Q.; Sun, J.; et al. Informedia @ TRECVID 2016. Available online: http://www-nlpir.nist.gov/projects/tvpubs/tv16.papers/inf.pdf (accessed on 26 October 2017).