applied sciences

MDPI

Article

A Person Re-Identification Scheme Using Local Multiscale Feature Embedding with Dual Pyramids

Kwangho Song and Yoo-Sung Kim *

Department of Information and Communication Engineering, Inha University, Incheon 22212, Korea; crossofjc@gmail.com

* Correspondence: yskim@inha.ac.kr; Tel.: +82-32-860-7450

Abstract: In this paper, we propose a new person re-identification scheme that uses dual pyramids to construct and utilize the local multiscale feature embedding that reflects different sizes and shapes of visual feature elements appearing in various areas of a person image. In the dual pyramids, a scale pyramid reflects the visual feature elements in various sizes and shapes, and a part pyramid selects elements and differently combines them for the feature embedding per each region of the person image. In the experiments, the performance of the cases with and without each pyramid were compared to verify that the proposed scheme has an optimal structure. The state-of-the-art studies known in the field of person re-identification were also compared for accuracy. According to the experimental results, the method proposed in this study showed a maximum of 99.25% Rank-1 accuracy according to the dataset used in the experiments. Based on the same dataset, the accuracy was determined to be about 3.55% higher than the previous studies, which used only person images, and about 1.25% higher than the other studies using additional meta-information besides images of persons.

Keywords: person re-identification; dual pyramid structure; multiscale feature; part-wise feature; high rank-1 accuracy

Citation: Song, K.; Kim, Y.-S. A Person Re-Identification Scheme Using Local Multiscale Feature Embedding with Dual Pyramids. *Appl. Sci.* **2021**, *11*, 3363. https://doi.org/10.3390/app11083363

Academic Editor: Hugo Pedro Proença

Received: 16 February 2021
Accepted: 2 April 2021
Published: 8 April 2021

1. Introduction

Person re-identification refers to identifying a pedestrian of interest based on external information or characteristics of walking from a number of people captured in single or multi-camera environments. It is being regarded as an essential technique for the intelligent video surveillance systems such as tracking offenders or searching for missing persons [1,2]. Recent person re-identification methods use deep neural network to transform the person images of both a query and a gallery set into feature embedding. Then, the similarity between the feature embedding of the query and the ones in the gallery set is estimated to verify the identity of person of interest by selecting similar images from the gallery set.

However, in an actual environment, the identity of a person can be easily confused depending on the differences in the captured time, place, or viewpoint. Because a large difference in appearance of person such as an appearing color, pattern of clothing, belongings, or exposed skin such as face, arms, and legs can be caused by the change of such conditions. Therefore, an intra-class variation that phenomenon of identifying the same person differently or an inter-class variation that phenomenon of identifying the different person as the same can happen easily. Due to such variations, the similarity between the same person can be estimated low, or the similarity between the different person may show high.

As a method to solve such variations, recent studies [3–12] have mainly used the following two types of methods. In the first method, a neural network uses additional meta-information such as time, place, and viewpoint at the time of photographing of the pedestrian with a person image [3,4]. Although this method has the advantage of improving the person re-identification accuracy, it is difficult to obtain such meta-information

automatically from an actual environment. Moreover, additional considerations and resources are required to process acquired information into a data form that can be input into a neural network. Therefore, the second method, which only uses images as input data, has been mainly studied. To obtain the performance close to the first method without additional meta-information, this method utilizes various preprocessing techniques or utilizes a number of auxiliary neural networks or special modules to extract more discernable representative features from the input image while constructing a neural network [5–12]. Although this method has an advantage of achieving performance close to the first method, the model can become excessively complicated due to the requirement of several auxiliary neural networks within the neural network or an algorithm in addition to the neural network [5,6,9]. Due to the limitations of the image segmentation method or filter shape used by the existing method, it is difficult to reflect the visual elements of various sizes and shapes appearing in the person image into the feature embedding [5–12].

Therefore, in order to enable robust re-identification against the intra/inter-class variations using only the given person images without any other meta-information, a new person re-identification method with dual pyramids that extracts various visual feature elements scattered in various areas of a person image and reflects them in the feature embedding is proposed in this paper. In dual pyramids, a scale pyramid reflects the visual feature elements in various sizes and shapes, and then a part pyramid selects visual feature elements and differently combines them for the feature embedding per each region of the person image. For such purpose, a scale pyramid has different sized kernels that are arranged in serial and parallel fashions, and a part pyramid uses divided and combined feature maps extracted from the scale pyramid to compose feature embedding. Moreover, the proposed model can provide accurate re-identification results using multi-scale features from various regions on the input image.

This paper is structured as follows. Section 2 introduces the existing studies that performed person re-identification based only on images of persons as related studies. In Section 3, the novel method for person re-identification proposed in this study is described. The feature embedding extracted from the neural network uses a dual pyramid structure to reflect various sizes and shapes of the visual feature elements shown in the image in the suggested method. Section 4 shows the experimental results related to each module and corresponding combinations to verify the structural validity of the proposed model and also describes the accuracy of the proposed person re-identification scheme with the publicly available data sets. Section 5 describes the qualitative and quantitative comparisons of the proposed scheme with other state-of-the-art methods. Finally, the conclusion of this paper and future research is described in Section 6.

2. Related Work

Existing methods that perform re-identification based only on person images [5–12] use neural networks with unique structures to construct more discriminating feature embedding by extracting as much information as possible from the given image. These methods can be classified into two broad categories depending on how the neural network selects the region of an image reflected in the feature embedding. The first method constitutes feature embedding that reflects all visual feature elements that exist in the entire image [5–8], and such feature embedding is called the global feature embedding [5–8]. Conversely, when multiple feature embedding is formed based on the visual elements of corresponding small area images and the small area images are considered as a combination of multiple area images when given an image, the embedding is referred to as local feature embedding [9–12].

First, the neural networks of the studies using global feature embedding [5–8] mostly focus on the generation of discriminative feature embedding, which denotes visual elements that are randomly distributed throughout the image. For example, in [5,6], the input person images were reconstructed into different sizes. For example, variously sized copies with different resolutions were initially made [5], or small patch images with various

sizes were separated from the pre-determined points of the foreground [6]. Subsequently, the reconstructed images were provided to the sub-networks that accept inputs suitable for differently sized images and then used to form feature embedding representing the corresponding person. This method has an advantage in that a neural network can extract and utilize a large amount of information from images of various sizes created based on a person image. This leads to exhibition of excellent performance without separate meta-information such as shooting time and location. However, as the sizes of the images used as inputs vary, several neural networks are required to accommodate images of different sizes. For example, two or more InceptionV3 networks [13] with 23.5 million in weight were used in [5], and 21 self-designed affiliated neural networks were used in [6], leading to excessively complex neural networks. In addition, adjusting previously determined resolution [5] or the separation point of the patches [6] becomes inevitable when applied to an actual environment for optimal performance. On the other hand, in [7,8], only one image was used as an input to the single neural network. Instead of any other preprocessing or auxiliary network, in [7] a unique module composed of a series of layer sets that consisted of a parallel connection of a series of one or two convolution layers was used, and in [8] one to four serially connected convolutional layers with kernels sized 3*3 each and arranged in parallel were placed within the neural network to provide a square-shaped receptive field of various sizes and reflect visual feature elements in the feature embedding. However, the important visual feature elements that can be used in identifying the person such as the person's arms, legs, and feet, clothing or patterns on it, and the person's belongings are mostly shown as rectangular shapes at various regions of the input image. Therefore, when the receptive fields of a neural network that can be given to the input image are limited to square shapes [7,8], properly reflecting visual elements of other shapes such as rectangular shapes with either longer horizontal or vertical direction can be problematic in the input embedding.

Previous studies using the local feature embedding also applied auxiliary network or similar techniques to find areas of body parts as a basic position for feature embedding extraction such as arms, legs, and torso [9]. However, this method was problematic because the performance of the person re-identification could be greatly varied depending on the body part detection result. Therefore, recent studies [10–12] simplified the meaning of a local area as a patch of a horizontally divided person image. The feature map extracted from the backbone such as ResNet50 [14] or VGG16 [15] is simply divided horizontally, and each region becomes an origin area of each local feature embedding. Accordingly, previous study [10] divides the feature map from ResNet50 [14] backbone into six regions horizontally and construct local feature embeddings representing each region by global average pooling. When each feature embedding is configured on the divided regions with fixed size, the visual feature elements that exist over two or more neighboring regions that are larger than the fixed size are difficult to be reflected in the feature embedding. Therefore, other studies [11,12] tried to solve this problem using a so-called feature pyramid. In [11,12], a feature pyramid was configured by combining six horizontally divided basic regions first, and then overlapping and tying between the successive neighboring regions, with a total of one to six in each group, creating a total of 21 combined regions. The study [12] also used a similar feature pyramid, which is configured by combining horizontally divided eight basic regions first, and then tying between the adjacent neighboring regions without overlapping, with a total of eight, four, two, and one in each group to create a total of 15 combined regions. These combined regions are converted into local feature embedding through the adding up of vectors created by global average pooling and global max pooling. However, in the case of constructing a pyramid with multiple combination of regions that overlap each other, each basic region is used to configure multiple combination of the regions. In particular, the basic region located in the center is used more often than the basic regions outside the center region. For this reason, regardless of the actual importance of each basic region, more regions located in the center of the image are more often unconditionally utilized to configure the combination of regions. This is a problem since the crucial visual

feature elements such as shoes, hair color, and exposed facial features are difficult to use as important sources. In addition, when the sum of the maximum pooling and the average pooling is used for the construction of the feature embedding, the value in the embedding eventually becomes an ambiguous value rather than an average or maximum, which may result in a deterioration in the discrimination of embedding.

Therefore, this study proposes a person re-identification method that applies a scale pyramid module that allows the neural network to extract visual feature elements that appear in various sizes and shapes in the input person image and a part pyramid module that allows configuration of more accurate feature embedding by appropriately reflecting extracting features according to each region. The scale pyramid module used in the proposed method is designed by connecting the convolutional layers with kernels of different sizes and shapes in series and a parallel manner, so the receptive fields of various sizes including square and rectangular shapes can be acquired from the input image. In addition, to construct discriminative feature embedding that evenly reflects visual feature elements that exist across either narrow or wide areas within the input image, a part pyramid module is horizontally divided into eight segments and then serially connected by regions, a total of eight, four, two, and one in each group, without a mutual overlap to produce a total of 15 regions to configure feature embedding in each region. Each region is converted into local feature embedding by using only global average pooling instead of the complex summation of maximum pooling and average pooling. This proposed method can form the feature embedding that independently reflects the visual feature elements of various sizes and shapes that exist on each region of the input person image. Based on this, preferable person re-identification without additional information on the person can be achieved.

3. Method

3.1. Overall Procedure of the Proposed Person Re-Identification Scheme Using Dual Pyramids

In this section, we introduce an overall procedure of the proposed person re-identification scheme using multiscale feature embedding made by dual pyramids, a scale pyramid and a part pyramid. As shown in the training process of (Figure 1), the proposed scheme uses only the person's image as an input without additional meta-information such as the captured time, location, and viewpoint. The input person image is converted into 15 part-level local feature embedding, and each embedding is used to train by the classification method that infers the person's ID with given cross entropy loss function as the ground truth (GT). The details of embedding, such as creation process and network architecture of sub-module are explained in the following Sections 3.2 and 3.3 with concrete examples.

In the subsequent inference process, in addition to the query image that requires re-identification, the gallery images to be used to verify the identity of the query image are given, and the pre-trained re-identification model is used to convert the query image and gallery images into the part-level feature embedding. Fifteen part-level feature embedding created from each image are connected to a single feature embedding, which designates multiscale features extracted from various locations of the input image and then uses them for similarity comparison. The multiscale feature embedding created from the query image is called query embedding, and that from the gallery images is called gallery embedding. The similarity comparison between feature embedding is made based on the cosine similarity, and the set of gallery embedding is sorted based on the similarity with the query embedding. In turn, the identity of the gallery image with the highest similarity is specified as the identity of the query image.

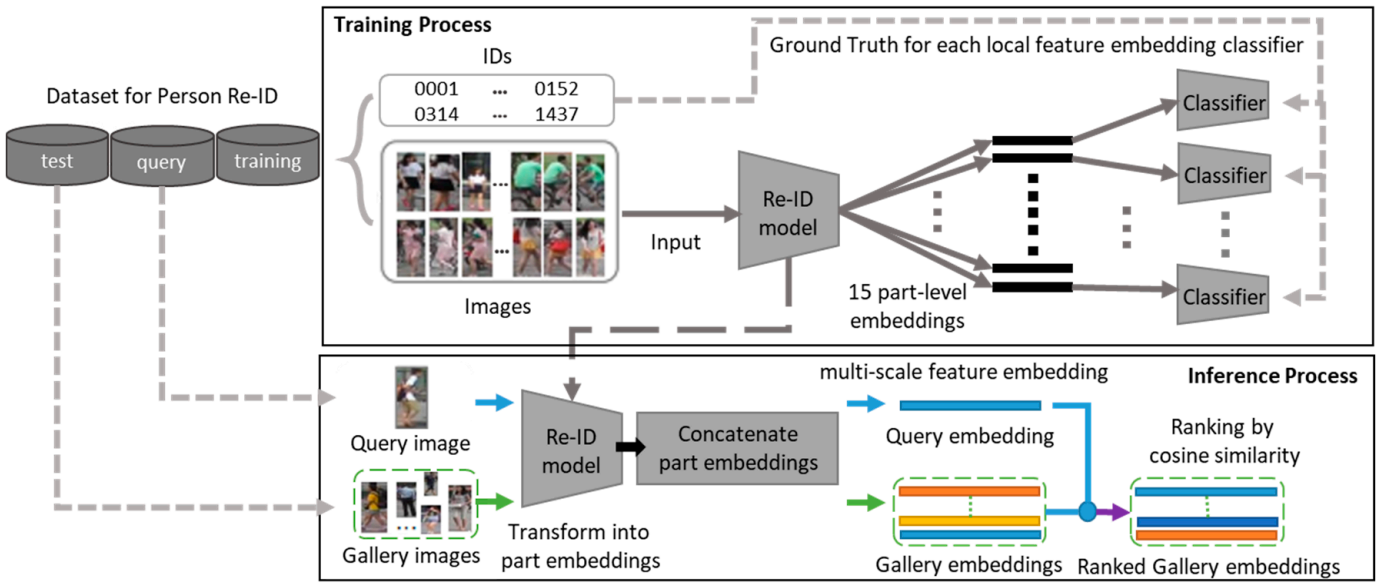

Figure 1. Process diagram of the proposed person re-identification model.

The structure of the neural network constituting the person re-identification model is configured in the order of the input, backbone network, scale pyramid, part pyramid, and output embedding module as shown in (Figure 2). First, as for the input image of the neural network, the rectangular images of 1:3 ratio with width of 128 and height of 384 are accepted instead of the square shapes commonly used by the image handling neural networks to reflect general body characteristic of longer height to width of the shoulder. The input image is transferred to the backbone and the subsequent convolutional layer to be converted into a feature map sized (8, 24, 512). Following the convolutional layer, batch normalization (BN) [16] and dropout [17] are placed to prevent overfitting by the previous layers. At this point, if the entire network of ResNet50 [14] is used as a backbone, the size of output is reduced to 1/32 of the input, which in turn limits the available features of the subsequent blocks. Therefore, the last convolution block in ResNet50 [14] is excised to produce a lower reduction ratio of 1/16 for more feature utilization by the subsequent layers. The extracted feature map from the backbone and subsequent convolution layer is delivered to the scale pyramid to produce six outputs, each of which is (8, 24, 512) size and represents the multiscale features from large to small and both square and rectangular-shaped visual elements of the input image. And these six output feature maps from the scale pyramid are transferred to the input of the part pyramid to generate 15 part-level local feature vectors sized (1, 1, 512*6) that originate from various regions of the input with different configuration as shown in (Figure 2). Finally, the vectors are converted to the local feature embedding, with a length of (512), by passing through a convolution layer with the last kernel sized 1*1, batch normalization (BN) layer, and dropout layer as shown in (Figure 2). The proposed person re-identification model can create the local feature embedding that reflects the visual feature elements of various sizes and shapes of each of the 15 regions created by the pyramid, and through this, the model can even utilize the detailed visual feature elements distributed throughout an image during the person re-identification process.

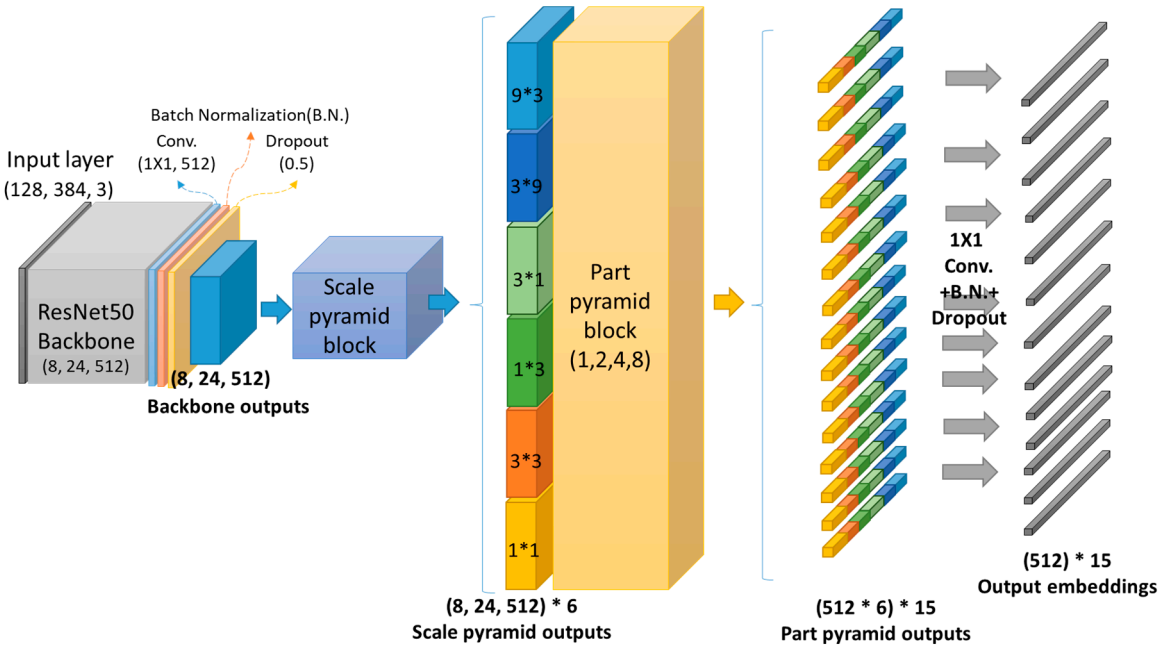

Figure 2. Network architecture of the proposed person re-identification model.

3.2. Scale Pyramid to Extract Multiscale Features

In this section, we introduce the scale pyramid in more detail. This pyramid can convert the visual elements of various sizes and shapes in the input image into the features used to compose the feature embedding.

The scale pyramid allows the person re-identification model to have receptive fields of various sizes including the rectangular shapes in addition to the square ones. The scale pyramid has a feature map sized (W, H, C) of (width, height, channel) as an input which is made by previous steps in the person re-identification model. It produces 6 outputs sized (W, H, C) from the receptive fields of different sizes and shapes as shown in (Figure 3). Looking first at the neural networks leading to the outputs Output_1 and Output_2, they share two consecutive convolutional layers with the kernels sized 1*1 and 3*3. Accordingly, the output feature map up to this point has a receptive field of 3*3 in size with respect to the input feature map, and since the two convolutional layers connected in parallel have the kernels sized 1*3 and 3*1, the receptive field of the output with vertically long rectangular shape (Output_1) of 3*9 and horizontally long rectangular shape (Output_2) of 9*3 can be accommodated with respect to the input feature map. The neural networks leading to the following Output_3 and Output_4 also share a convolutional layer with a kernel sized 1*1, and since the two convolutional layers connected in parallel have kernels sized 1*3 and 3*1, the receptive field of the output acquires a vertically long rectangular shape sized 1*3 (Output_3) and a horizontally long rectangular shape sized 3*1 (Output_4) with respect to the input feature map. Unlike the other parts, the neural network leading to Output_5 secures a square-shaped receptive field sized 3*3 through average pooling and learns the features of the corresponding region through a convolution layer with a kernel sized 1*1. Finally, the neural network leading to Output_6 uses a convolutional layer with a kernel sized 1*1 to secure the smallest 1*1 sized square-shaped receptive field.

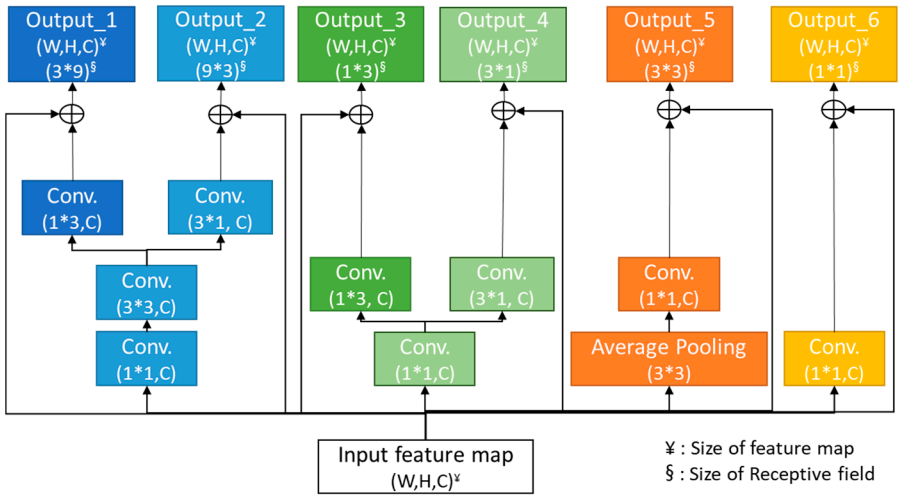

Figure 3. Network architecture of the scale pyramid.

Therefore, the scale pyramid can secure the receptive fields of 6 different sizes and shapes that include the square shapes sized 3*3 and 1*1 as well as the rectangular shapes sized 3*9, 9*3, 1*3, and 3*1 with respect to each arbitrary point of the (8, 24, 512) sized feature map, which originated from the output feature map of the backbone as shown in the example of (Figure 4). Regarding the input image, the scale pyramid can extract large and small, square and rectangular-shaped visual elements for all regions of the image as the features to reflect in the embedding. Accordingly, the proposed person re-identification model can utilize the visual feature elements of different sizes and shapes that exist at various locations on the input image during the formation of the feature embedding.

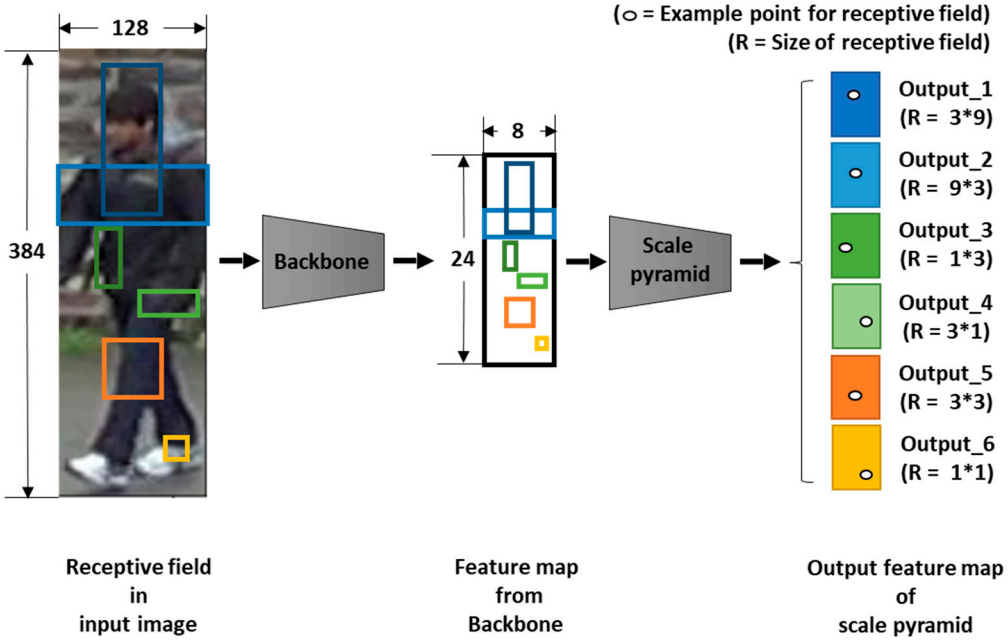

Figure 4. Receptive field example of scale pyramid on input image and feature map.

3.3. Part Pyramid to Generate Local Feature Embedding in Part-Level

In this section, we introduce the part pyramid, which can convert part-level information in the input image, which originates from various regions with different configuration, into the local feature maps.

As shown in (Figure 5), the architecture of the proposed part pyramid is similar to the feature pyramid in [12] in that it receives one input feature map and constructs 15 output regions. However, as described below, there are differences in the size of input/output feature maps and the method of configuring individual regions, and in the method of global pooling for each region. First, 6 feature maps sized (W, H, C) that are produced from the scale pyramid are transferred to the input of the part pyramid block as shown in (Figure 5). The part pyramid block reconstructs the received feature maps to 15 different feature maps, each of which represents a respective field that has one of the four different sizes to organize the local feature embedding evenly with the visual feature elements that appear in the various areas of the input image such as patterns or color of bags, belongings, or clothes as well as the narrow area of the input image such as shoes or hair color without bias of a particular region. For this purpose, the part pyramid block is constructed from one feature map by concatenating 6 feature maps that are received as inputs and horizontally dividing the maps into 8 regions to form basic local feature maps from Base_1 to Base_8 with a size of (W, H/8, C*6) as shown in (Figure 5). Subsequently, to extract the visual feature elements distributed over a wider area, 4 types of 15 combination local feature maps are configured with 1 combination feature map connected with the Output_1_1 sized (W, H, C*6), which is composed of 8 neighboring basic local feature maps without overlapping, 2 combination feature maps connected with the Output_2_1 and Output_2_2 sized (W, H/2, C*6) that are attached by 4 neighboring maps, and 4 combination feature maps connected with Output_4_1, 2, 3 and 4 sized (W, H/4, C*6) that are attached by 2 neighboring maps, and 8 feature maps (from Output_8_1 to Output_8_8) that are directly from each of the 8 basic local feature maps. These 15 local feature maps created through a series of processes are converted into local feature vectors sized (C*6) by the global average pooling only to avoid the ambiguity of feature value in the vector.

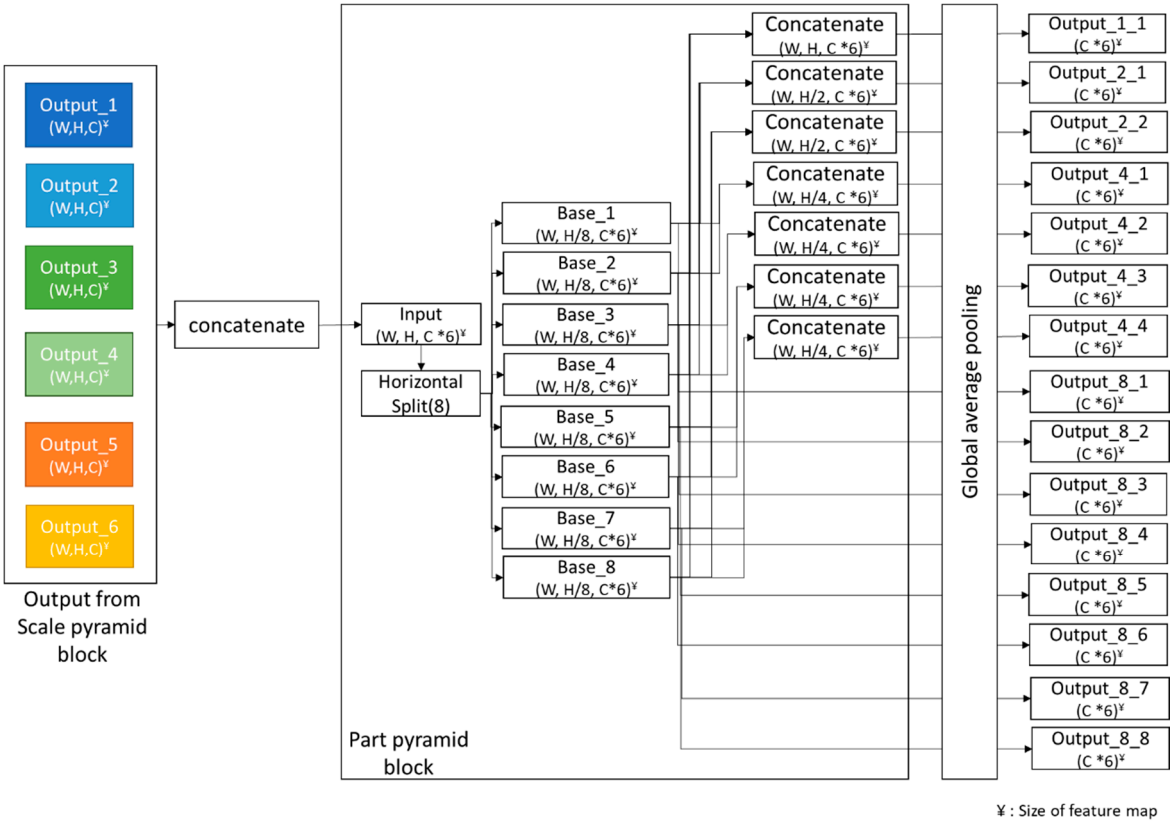

Figure 5. Network architecture of part pyramid.

Through this, the model can create the part-level local feature embedding originating from various regions of different configurations. As shown by the person in (Figure 6), the feature embedding extracted from 8 basic regions reflects the visual feature elements that appear in order of the hair color, appearance of the face, and characteristics of clothing or belongings on the chest and abdomen, characteristics of clothing or belongings on the buttocks, characteristics of clothing on the thighs and calves, and shoe color. In addition, 4 feature maps that gather 2 basic local areas each can extract feature embedding that appears in order of the hair color and length and appearance of face, overall characteristics of the top, characteristics of the top section of the bottom, and characteristics of the lower section of the bottom and shoes. Two feature maps that gathered 4 basic local areas each can extract feature embedding that appears in order of the overall characteristics of the top and overall characteristics of the bottom. Lastly, 1 feature map that gathers 8 basic local areas each can extract feature embedding of the overall visual feature element of the person image. Therefore, the 15 local feature maps created by the part pyramid accommodate regions corresponding to different ranges on the input image to configure feature embedding per region to reflect unique characteristics of each region. A preventive measure for biased training of a particular region by the feature embedding and configuration of excessively large number of local embedding at the same time was performed by preventing overlapping between the regions during the configuration of each region.

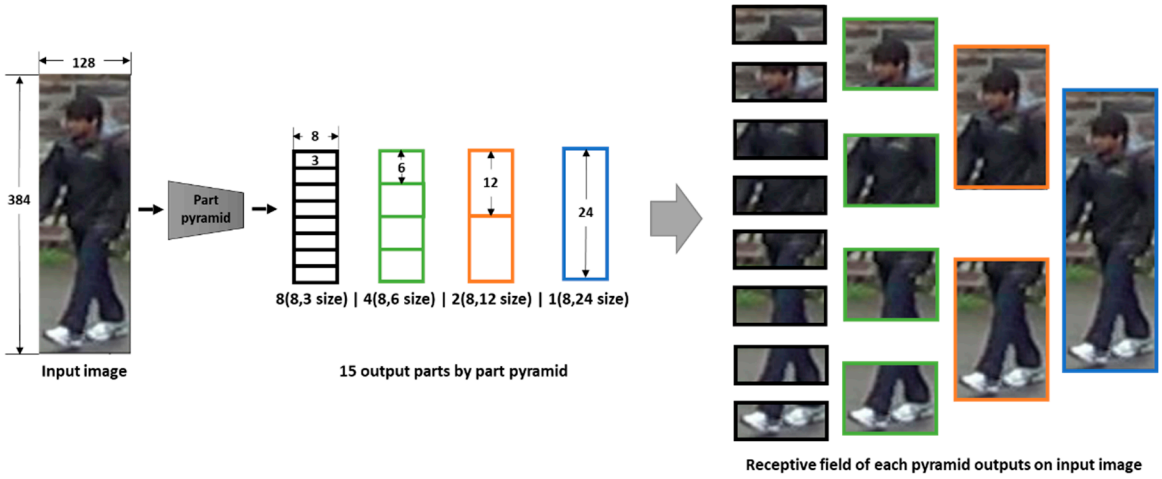

Figure 6. Receptive field example of part pyramid on input image.

4. Experiment and Analysis

The dataset of DukeMTMC-reID [18] which is widely used for the development of person re-identification models, was used for this study. The DukeMTMC-reID [18] dataset composed of the images of 702 persons include 16,522 training images, 17,661 gallery images, and 2228 query images. In addition, the Market-1501 [19] dataset, which is widely used in the field of person re-identification, was used in the experiment to evaluate the performance of the model and increase the evaluation objectivity by examining the dependency of the model on a specific dataset. The Market-1501 [19] dataset, which is composed of images of 750 persons, consists of 12,936 learning images, 13,115 gallery images, and 3368 query images.

Prior to the model training, all training images went through the data augmentation process consisting of horizontal flip, random erase, and normalization, and accordingly, 66,088 images for DukeMTMC-reID [18] and 51,744 images for Market-1501 [19] were used during actual training. As the ground truth label for each training image, one-hot encoding vector with corresponding ID value set at 1 for the person on a given image and 0 for all others was used in general. The Adam optimization function [20] was also used during a total of 80 epochs. At this time, the starting training rate was set at 0.001 and was reduced

by 0.1-fold for every 30 epochs elapsed. In addition, as the validation data that verified the overfitting status during the training process, randomly selected 6609 images consisting of 10 percent of all training data were utilized. If the calculated validation loss based on the validation data was the smallest, it was saved and used as the weight of the model. Moreover, the zero paddings that filled the remaining part with zero value were set to allow for the inputs during the usage of the kernels by all convolutional layers, and the ReLU [21] function was allowed for all functions to activate the outputs of the layers for the neural network construction.

As a method to evaluate the performance of the model, the Rank-1 accuracy which decides on the equivalency based on the query image ID and first rank among the gallery embedding that are aligned based on the cosine similarity was used. Because of the special characteristic of the person re-identification among the intelligent video security schemes that require accuracy in identification such as tracking abnormal behaviors or searching for missing persons, the superior accuracy of the Rank-1 identification result is a critical performance indicator of the proposed method.

The first experiment was conducted by switching among ResNet50 [14], VGG 16 [15], and SeResnet101 [22], which were used in by previous person re-identification experiments as the backbone neural network to set the optimal structure of the neural network constituting the re-identification model.

In the experimental results shown in (Table 1), ResNet50 [14] showed the best performance among the three backbones. The shallow neural network VGG16 [15] showed difficulty in constructing the feature embedding that sufficiently reflected the visual elements and discriminated person ID with the neural network. On the other hand, the excessively deep neural network SeResNet101 [22] showed decreased performance since unnecessary visual elements as well as the valid visual elements were reflected in the feature embedding.

Table 1. Configuration evaluation result for backbone module.

Training Setting for Backbone with ImageNet [23] Pre-Trained Weight	Rank-1
ResNet50 [14]	**84.51%**
SeResNet101 [22]	82.58%
VGG16 [15]	70.51%

In the second experiment, performance evaluation was conducted by applying different sized pyramids with five configurations composed of the outputs of variously sized receptive fields that included 1*1, 3*3, and 9*9 sized square shapes and 1*3, 3*1, 9*3, and 3*9 sized rectangular shapes on the output feature maps extracted from a backbone as suggested in (Table 2). In (Table 2), the receptive field included in each configuration is marked with "O", and the receptive field that is not included is marked with "X".

Table 2. Configuration evaluation result for scale pyramid block.

Training Setting for Scale Pyramid with ResNet50 [14] Backbone							Rank-1
Receptive Field Size of Scale Pyramid Outputs							
1*1	3*3	9*9	1*3	3*1	9*3	3*9	
O	O	X	X	X	X	X	92.77%
O	O	O	X	X	X	X	90.88%
O	O	X	O	O	X	X	92.86%
O	**O**	**X**	**O**	**O**	**O**	**O**	**92.99%**
O	O	O	O	O	O	O	92.36%

As the results of the experiment, better performance was shown for the cases with both shaped receptive fields instead of the cases with only the square-shaped receptive fields

sized (1*1, 3*3) or (1*1, 3*3, 9*9). Moreover, the performance increased with the increased number of available receptive fields and showed optimal performance in the configurations with six outputs sized (1*1, 3*3, 1*3, 3*1, 3*9, 9*3). However, the performance decreased whenever a pyramid branch with a receptive field sized (9*9) was added. It is estimated that a receptive field sized (9*9) with the input feature map sized (8, 24) may be too large to extract local features properly. Through this experiment, it was shown that having shapes of receptive fields close to the visual feature elements that appear in the person image in the neural network in addition to square-shaped receptive fields commonly used in neural networks is more effective in constructing more discerning feature embedding.

In the third experiment, the performance evaluation experiment was conducted on the 15 feature vectors extracted from the part pyramid based on six feature maps from the scale pyramid by changing the composition of the global pooling mechanism for local feature embedding that constitutes the output of part pyramid as suggested in (Table 3).

Table 3. Configuration evaluation result for part pyramid block.

Training Setting for 15 Outputs of Part Pyramid with 6 Output Scale Pyramids		Rank-1
Global Pooling Step in Part Pyramid		
Global Average Pooling	Global Max Pooling	
Used	**Not used**	**94.79%**
Not used	Used	93.49%
Used	Used	92.72%

As the results of the experiment, better performance was shown for the case with only the global average pooling used compared to only the global max pooling used or the sum of global average pooling and global max pooling used. Especially, the lowest performance was produced when the sum of the global max pooling and global average pooling was used. This can be interpreted as meaning that when these values are used together, the value becomes ambiguous, and the performance may decrease compared to when using each alone.

Therefore, as shown in the previous experiments, the model composed of the scale pyramid with six outputs sized (1*1, 3*3, 1*3, 3*1, 3*9, 9*3) based on the backbone of ResNet50 [14] and the part pyramid with 15 outputs of 0, 2, 4, and 8 divisions with global average pooling can be said to have the optimal configurations for the person re-identification based on the method proposed in this study. Considering the weight constituting this model is 43 million and the weight of neural networks used as the backbone is 47 million in the model of [5], the number of auxiliary neural networks is 21 for the model of [6], and the model of [10] uses all of ResNet50 [14] as the backbone and more complex composite local feature embedding than current study, the complexity of the neural networks of the proposed method is considered to be less than or similar to previous studies.

As the final experiment, the proposed model was trained by the datasets of DukeMTMC-reID [18] and Market-1501 [19], and the performance of the proposed model was evaluated and showed in (Table 4).

Table 4. Performance evaluation using DukeMTMC-reID [18] and Market-1501 [19] datasets.

Training and Test Dataset	Rank-1
DukeMTMC-reID [18]	94.79%
Market1501 [19]	99.25%

As the results, the proposed model showed the best performance in terms of the Rank-1 accuracy of 94.79% based on the DukeMTMC-reID [18] dataset. The excellent performance

of 99.25% in terms of the Rank-1 accuracy was also observed with the Market-1501 [19] dataset. The results suggest that the proposed method can show a robust performance regardless of the shooting environments or the type of person composition of the dataset used for training and evaluation.

5. Discussion

The main aim of this study was to develop a dual pyramid structure with more discriminative feature embedding for a person re-identification. To accurately detect the gallery that is the same as the query, discriminative feature embedding was used, which is the connection of 15 local feature embedding by the proposed scheme using scale and part pyramid structure. According to the analysis of the experimental results in (Table 5), the performance rates of the proposed scheme based on Rank-1 accuracy are 94.79% and 99.25% for the DukeMTMC-reID [18] dataset and the Market-1501 [19] dataset, respectively.

Table 5. Performance comparison with recent studies [3–12].

Model	Rank-1		Additional Meta-Information
	DukeMTMC [18]	Market-1501 [19]	
DL multi-scale Representations [5]	79.2%	88.9%	Not used
PCB [10]	83.3%	92.3%	Not used
Horizontal Pyramid Matching [12]	86.6%	94.2%	Not used
OSNet [8]	88.6%	94.8%	Not used
Pyramidal Person Re-Id [11]	89.0%	95.7%	Not used
Viewpoint-Aware Loss [4]	91.61%	96.79%	View-point info.
St-reID [3]	94.00%	98.00%	Spatio-temporal info.
Ours	**94.79%**	**99.25%**	Not used

In recent years, many studies have reported high performance in person re-identification using various deep learning approaches [3–12]. Among the studies, some used additional meta-information [3,4], multi-scale features [5,8], or localized features [10–12] as useful elements to determine the identity of the person in the image.

To compare the performance results of these approaches, the approaches of [3] and [4] with the Rank-1 accuracy rate of 94.0% and 91.61% based on the DukeMTMC-reID [18] dataset and the Rank-1 accuracy rate of 98.00% and 96.79% based on the Market-1501 [19] dataset, which is a lower performance than proposed scheme, are shown. These results suggest that only the person image can be used for re-identification without additional meta-information such as shooting time, location, and viewpoint for a comparatively superior performance. The approaches of [5] and [8] with the Rank-1 accuracy rate of 79.2% and 88.6% based on the DukeMTMC-reID [18] dataset and the Rank-1 accuracy rate of 88.9% and 94.8% based on the Market-1501 [19] dataset are also shown. Moreover, the approaches of [10], [11], and [12] with the Rank-1 accuracy rate of 83.3%, 89.0%, and 86.6% based on the DukeMTMC-reID [18] dataset and the Rank-1 accuracy rate of 92.3%, 95.7%, and 94.2% based on the Market-1501 [19] dataset are shown. All of them have a lower performance than the proposed scheme, which suggests that using multi-scale features and localized features together can make better performance than using them separately.

6. Conclusions and Future Work

In this study, a novel person re-identification method based on dual pyramids of a scale pyramid and a part pyramid was proposed for more accurate person re-identification results by extracting the visual feature elements of various sizes and shapes appearing in different regions of a person image. The scale pyramid applied to the proposed model enables more diverse and accurate feature extraction by allowing different sizes that include the rectangular-shaped feature receptive fields as well as the square-shaped feature receptive fields used in the previous studies with neural networks. The part pyramid allows various regions of an image to form innate feature embedding, thereby creating

feature embedding that reflects detailed visual feature elements that exist in each region for more accurate person re-identification. In addition, since the proposed method shows superior Rank-1 accuracy to the method using only the square-shaped features or the method using only global feature embedding, the multi-scaled regional features used by the dual pyramid structure are shown to be valuable in person re-identification.

In the future, in order to obtain more accurate re-identification results, research is planned to select relatively more important areas for each region using the attention mechanism during the extraction of regional multi-scale features.

Author Contributions: Conceptualization, K.S.; methodology, K.S.; software, K.S.; validation, K.S. and Y.-S.K.; investigation, K.S.; resources, K.S. and Y.-S.K.; writing, original draft preparation, K.S.; writing, review and editing, K.S. and Y.-S.K.; visualization, K.S.; supervision, Y.-S.K.; project administration, Y.-S.K. All authors read and agreed to the published version of the manuscript.

Funding: This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00203, Development of 5G-based Predictive Visual Security Technology for Preemptive Threat Response).

Conflicts of Interest: The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person re-identification: Past, present and future. *arXiv* **2016**, arXiv:1610.02984.
2. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S. Deep Learning for Person Re-identification: A Survey and Outlook. *arXiv* **2020**, arXiv:2001.04193v1.
3. Wang, G.; Lai, J.; Huang, P.; Xie, X. Spatial-Temporal Person Re-identification. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19), Honolulu, HI, USA, 27 January–1 February 2019; pp. 8933–8940.
4. Zhihui, Z.; Xinyang, J.; Feng, Z.; Xiaowei, G.; Feiyue, H.; Weishi, Z.; Xing, S. Viewpoint-Aware Loss with Angular Regularization for Person Re-Identification. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19), Honolulu, HI, USA, 27 January–1 February 2019.
5. Yanbei, C.; Xiatian, Z.; Shaogang, G. Person Re-identification by Deep Learning Multi-scale Representations. In Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW), Venice, Italy, 22–29 October 2017; pp. 2590–2600.
6. XiuJie, Y.; Ping, C. Person re-identification based on multi-scale convolutional network. *Multimed. Tools Appl.* **2020**, *79*, 9299–9313.
7. Qian, X.; Fu, Y.; Jiang, Y.-G.; Xiang, T.; Xue, X. Multi-scale deep learning architectures for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5399–5408.
8. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-Scale Feature Learning for Person Re-Identification. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 3702–3712.
9. Zhao, G.; Jiang, J.; Liu, J.; Yu, Y.; Wen, J. Improving Person Re-identification by Body Parts Segmentation Generated by GAN. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.
10. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 480–496.
11. Zheng, F.; Deng, C. Pyramidal Person Re-Identification via Multi-Dynamic training. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
12. Fu, Y.; Wei, Y.; Zhou, Y.; Shi, H.; Huang, G.; Wang, X.; Yao, Z.; Huang, T. Horizontal Pyramid Matching for Person Re-Identification. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19), Honolulu, HI, USA, 27 January–1 February 2019; pp. 8295–8302.
13. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
15. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
16. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 448–456.
17. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.

18. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In Proceedings of the European Conference on Computer Vision (ECCV) Workshop, Amsterdam, The Netherlands, 8–10 October 2016.
19. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 11–18 December 2015; pp. 1116–1124.
20. Kingma, D.; Ba, J.A. A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
21. Agarap, A. Deep Learning using Rectified Linear Units (ReLU). *arXiv* **2018**, arXiv:1803.08375.
22. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *arXiv* **2019**, arXiv:1709.01507v4. [CrossRef] [PubMed]
23. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255.