Spatial and Temporal Segmented Dense Trajectories for Gesture Recognition

Kaho Yamada*[a], Takeshi Yoshida[b], Kazuhiko Sumi[b]
Hitoshi Habe[c], Ikuhisa Mitsugami[d]
[a]Graduate School of Science and Engineering, Aoyama Gakuin University/5-10-1 Fuchinobe, Chuo-ku, Sagamihara-shi, Kanagawa 252-5258, Japan; [b]College of Science and Engineering, Aoyama Gakuin University/5-10-1 Fuchinobe, Chuo-ku, Sagamihara-shi, Kanagawa 252-5258, Japan; [c]Faculty of Science and Engineering, Kindai University/3-4-1 Kowakae, Higashiosaka City, Osaka 577-8502, Japan; [d]Institute of Scientific and Industrial Research, Osaka University/8-1 Mihogaoka, Ibaraki City, Osaka 567-0047, Japan

ABSTRACT

Recently dense trajectories [1] have shown to be a successful video representation for action recognition, and have achieved state-of-the-art results on a variety of datasets. However, there are problems to recognize similar and fine-grained motion if we apply these trajectories to gesture recognition. In this paper, we propose new method in which dense trajectories are calculated in segmented regions around the detected human body parts. Spatial segmentation is achieved by body parts detection [2]. Temporal segmentation is done for every fixed number of video frames. The proposed method enables to remove background video noises. In addition, our method enables to recognize similar and fine-grained motion. Since only few video datasets are available for gesture classification, we also built new gesture dataset and evaluated our method with the dataset. Experimental results show that our method outperformed the original dense trajectories.

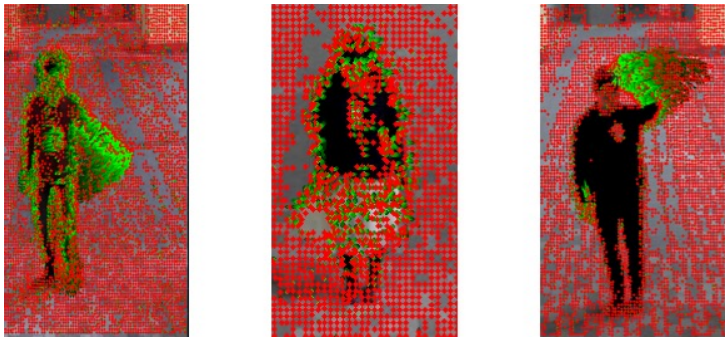
Keywords: Gesture recognition, Dense trajectories, Body part detection

1. INTRODUCTION

Recently, it is becoming popular providing information to each individual person in public space, such as stations and shopping malls. If the customers are in a group, it is preferable to optimize recommendations will be changed from the one for an individual person. However, a technology detecting group and estimating relationship have not been established yet. To estimate relationship, High-precision group detection is needed. Group detection has been studied by several researchers [3-7]. But, it is difficult to detect group only with features proposed by their method, such as the head pose and the distance between two pedestrians. If we can recognize human gesture, persons will be recognized accurately as a group. Therefore, action including interaction recognition, i.e. gesture recognition, is important to improve the precision of the group detection.

Action recognition has been studied from various perspectives [8-13]. Specially, trajectories around humans convey critical long-period information on human behaviors. Therefore, trajectory-based action recognition has been extensively studied in the past few years [1, 13, 15, 17, 18]. Matikainen et al. [13] extracted trajectories by using a standard Kanade-Lucas-Tomasi (KLT) tracker [14]. The trajectories were clustered. The elements of an affine transformation matrix which was computed for each cluster center were used to represent the trajectories. Sun et al. [15] extracted trajectories by matching SIFT descriptors [16] between two consecutive frames. To limit the effect of incorrect matches, they imposed a unique-match constraint among the descriptors and discarded matches that are too far apart. Sun et al. [17] combined KLT with SIFT trajectories to extract long-duration trajectories and to increase the trajectories density. To assure a dense coverage with trajectories, random points are sampled for tracking within the region of existing trajectories. Takahashi et al. [18] extracted fixed-dimensional features from KLT trajectories and SURF features. The SURF features were extracted by calculating the SURF descriptors [19] at the end point of the tracking. Among these trajectories, dense trajectories which proposed by Wang et al. [1] have shown to be an efficient video representation for action recognition, and have achieved state-of-the-art results on a variety of datasets. They densely sampled key-points for each frame, and tracked key-points in the video based on optical flow to obtain trajectories. They computed multiple descriptors, i.e. trajectory, Histogram of

Gradient (HOG) [20], Histogram of Optical Flow (HOF) [21] and Motion Boundary Histogram (MBH) [22] along the trajectories of key-points to capture shape, appearance and motion information. Figure 1 shows examples of dense trajectories. However, these trajectories are accumulated for whole image regions and times. Although it can classify major action recognition such as sitting and walking, it is not suitable for similar and fine-grained motion of human, such as pointing with hands, waving hands, and other gestures.



(a) (b) (c)

Figure 1. **Visualization of dense trajectories.** The red dots indicate the point positions in the current frame. The green lines indicate dense trajectories. (a): Visualization of dense trajectories for "Pointing with hands". (b) Visualization of dense trajectories for "Nodding". (c): Visualization of dense trajectories for "Waving hands".

In this paper, we improve dense trajectories by extracting local features with spatio-temporal segmentation. To segment spaces, we utilize a body part detection [2] and improve region range of each body part for each video frame. To segment times, we divide video frames into fixed time durations. These segmentations enable to remove video noises. In addition, they enable to recognize similar and fine-grained motion. Since only few video datasets are available for gesture classification, we also built our own dataset. We compared our method with the original dense trajectories on this dataset. This result showed that our method is state of the art for gesture recognition.

This paper is organized as follows. Section 2 presents our video representation with spatial and temporal dense trajectories. The experimental results are given in Section 3. Section 4 concludes the paper.

2. SPATIAL AND TEMPORAL SEGMENTED DENSE TRAJECTORIES

The proposed method has added two steps (Step 1 and 2) to the original dense trajectories which have three steps. Therefore, our gesture recognition method consists of five steps, which are shown in Figure 2. Step 1 is spatial segmentation. This step sets regions around the detected human body parts [2]. Step 2 is temporal segmentation. We divide video frames into fixed time durations. Step 3 is key-points detection and tracking. Step 4 is feature extraction. These Steps are based on the original dense trajectories [1]. Step 5 is classification. Bag-of-Features (BoF) approach [23] and Support Vector Machine (SVM) classifier [24] are used to recognize gesture. The following subsections explain each step.

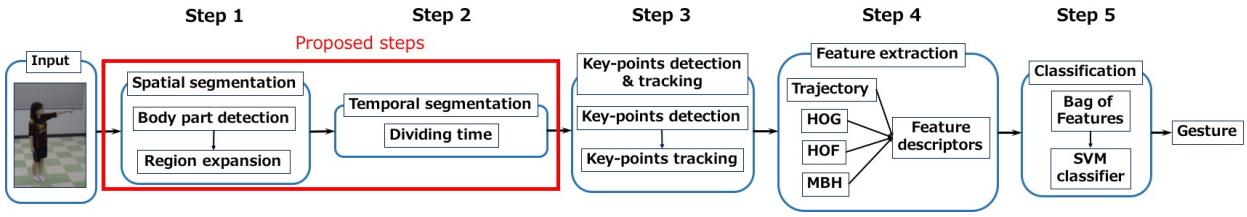


Figure 2. **Process of human gesture recognition.** The whole process is composed of five steps. Step 1 and 2 are the proposed steps. These steps make us extract local features.

2.1 Gesture dataset

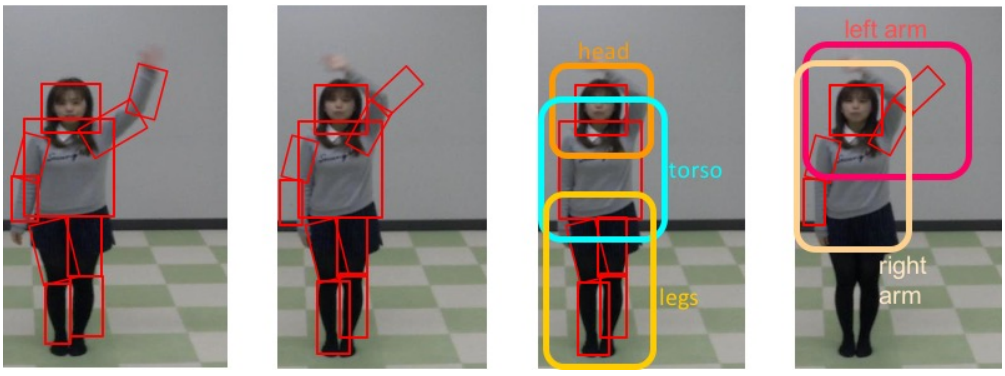
Among the datasets of videos released to date, only few video datasets are available for gesture classification. In this paper, we present new video dataset named gesture dataset. To select gesture categories, we manually counted the number of gesture used in the same group from school festival videos which captured guests enjoying the festival at corridor. Since Pointing with hands, Nodding and Waving hands are the dominant (70.8%) of gestures, we recognize these gestures. The dataset contains 456 videos in total from three gesture categories, with each category containing at least 100 videos (see Figure 3 for examples). There are 43 subjects wearing different clothes in two different environments, such as semi-outdoor and indoor. Each gesture class has 70 videos for training and 30 videos for testing.

(a) Pointing (b) Nodding (c) Waving (d) Pointing (e) Nodding (f) Waving

Figure 3. **Simple frames from the gesture dataset used in our experiments.** (a)-(c): Indoor shot. (d)-(f): Semi-outdoor shot.

2.2 Spatial segmentation

We set regions around each human body part i.e., head, torso, arms and legs for each video frame. For this purpose, we use a method inspired by Rothrock et al. [2]. They proposed a framework for human pose estimation using an articulated grammar model. Figure 4 (a) and (b) show the results of their method. These results present that accuracy of arms detection is lower than that of head, torso and legs detection (Figure 4 (b)). In addition, their predicted regions are narrower than the actual regions, as shown in Figure 4 (a) and (b). In this paper, we expand regions size around their predicted head, torso and legs regions by 30%, as shown in Figure 4 (c). Arms regions are also expanded with circumscribed rectangles size of their predicted head and arms by 30%, as shown in Figure 4 (d). We choose these regions size since it results in better performance, as discussed in Section 3.1. We calculate dense trajectories in each part region.



(a) (b) (c) (d)

Figure 4. **Spatial segmentation.** (a): A successful example of a method inspired by Rothrock et al. [2]. (b): A failure example of their method. (c): Regions of head, torso and legs. (d): Regions of arms

2.3 Temporal segmentation

To segment times, we divide video frames into fixed time durations as shown in Figure 5. In this paper, the frames are divided into four segments. We choose this number since it results in better performance, as discussed in Section 3.1.

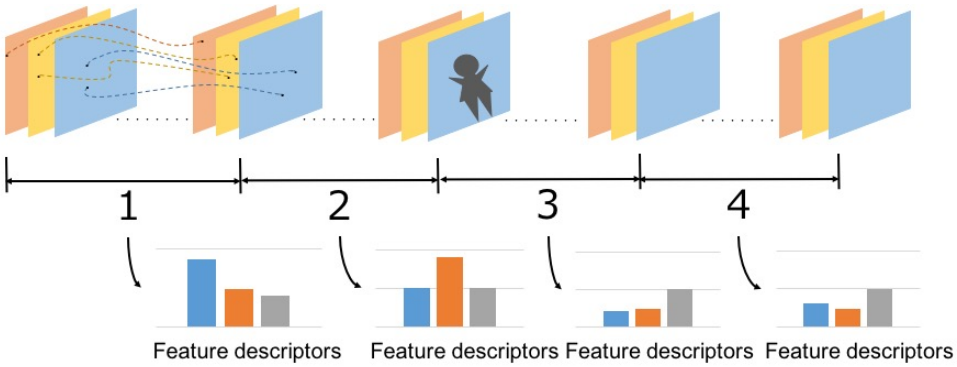


Figure 5. **Process of temporal segmentation.**

2.4 Key-points detection and tracking

Key-points are densely sampled on a grid spaced by five pixels in multi scales. We use eight spatial scales increased by a factor of $1/\sqrt{2}$. Most points in homogeneous areas are eliminated by a threshold for the smaller eigenvalue of their autocorrelation matrices. Then these sampled points are tracked by media filtering of dense flow field $\omega = (u_t, v_t)$ [25].

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega_t)|_{(\bar{x}_t, \bar{y}_t)}, \quad (1)$$

where M is the median filter kernel, $*$ is convolutional operation, and $(\bar{x}_t, \bar{y}_t)$ is the rounded position of (x_t, y_t). To avoid the drifting problem of tracking, we limit the length of a trajectory to 15 frames. Those static trajectories are removed as they lack motion information, and other trajectories with suddenly large displacement are also removed, since they are obviously incorrect due to inaccurate optical flow.

2.5 Feature extraction

For each trajectory, we compute several descriptors i.e., trajectory, Histogram of Gradient (HOG), Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH) with exactly the same parameters as [1]. The trajectory descriptor is a concatenation of normalized displacement vectors. The other descriptors are computed within the space-time volume aligned with the trajectory to encode the motion information, as shown in Figure 6. The size of the volume is 32×32 pixels and 15 frames long. To embed structure information, the volume is subdivided into a grid of size $2\times2\times3$. HOG is based on the orientation of image gradients and captures the static appearance information. Both HOF and MBH measure motion information, and are based on optical flow. HOF directly quantizes the orientation of flow vectors. MBH splits the optical flow into horizontal and vertical components, and quantizes the derivatives of each component. The final dimensions of these descriptors are 30 for trajectory, 96 for HOG, 108 for HOF and 192 for MBH. We extract these feature descriptors from each part region and time and combine them.

2.6 Classification

To encode features, we use BoF approach. We train a codebook for each descriptor type using 100,000 randomly sampled features with k–means. The size of the codebook is set to 4000. For classification, we use a non-linear SVM with RBF-χ^2 kernel [21] and different descriptor types are combined by summing their kernel matrices normalized by the average distance.

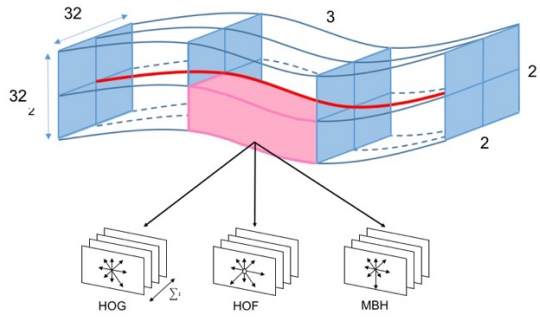


Figure 6. **Process of future extraction.** The trajectory is represented by relative point coordinates, and the descriptors are computed along the trajectory in a 32×32 pixels neiborhood, which is divided into 2×2×3 cells.

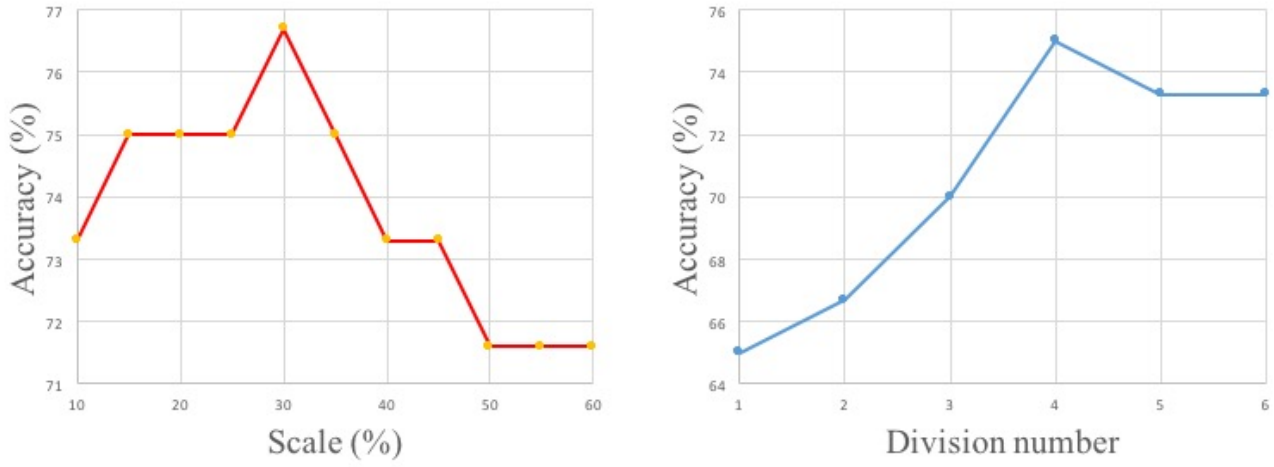
3. EXPERIMENTAL RESULTS

In this section, we first described parameter tuning in the proposed method. Then, we gave the experimental results and compared to the original dense trajectories [1], the spatial segmented dense trajectories and the temporal segmented dense trajectories on the gesture dataset.

3.1 Parameter Tuning

Region size. To specify size of detected human body parts regions for spatial segmentation, we explore different size of their regions on the gesture dataset by using spatial segmented dense trajectories. In this exploration experiment, the results are shown in Figure 7 (a). We vary the size from 10 to 60%. The results show that size 30% achieves the high performance. Thus, we apply the size as 30% in the remainder this section.

Division number. We also explore different division number on this dataset by extracting feature descriptors from temporal segmented dense trajectories, to specify division number for temporal segmentation. We vary the division number from 1 to 6. The results are shown in Figure 7 (b). The results show that division number four achieves the high performance. In this paper, we fix the division number as four.



(a) (b)

Figure 7. **Parameter tuning of different settings in the proposed method on the gesture dataset.** (a): Comparison of different size of body parts regions for spatial segmentation. (b): Comparison of different division number for temporal segmentation.

3.2 Comparison to the original dense trajectories

We compared our method (spatial and temporal segmented dense trajectories) to the original dense trajectories [1], the spatial segmented dense trajectories and the temporal segmented dense trajectories on the gesture dataset. We summarize the experimental results in Table 1. The confusion matrices for the original dense trajectories and our method is shown in Figure 8. Table 1 and Figure 8 show that our method outperformed the original dense trajectories on all categories. In the original dense trajectories, fine-grained motion such as "Nodding" tended to be wrongly recognized as other gesture categories (Figure 8 (a)) if there are background video noises. Then, "Pointing" tended to be recognized as "Waving" (Figure 8 (a)) because they include similar motion. However, our method and the spatial segmented dense trajectories for "Nodding" outperformed the original dense trajectories by 15%. From this result, the spatial segmentation enabled to remove video noises and recognize fine-granted motion. Our method for "Pointing" also outperformed the original dense trajectories by 30%. This result shows that spatial and temporal segmentation is effective for classifying similar motion (Figure 8 (b)).

However, our method for "Waving" show lower performance than the temporal segmented dense trajectories. The reason is that when the region of head overlaps with the region of arm, we fail human body parts detection and spatial segmentation of them, as shown Figure 9. To result in better performance, we will need a head detection which takes occlusion into account.

Table 1. **Gesture recognition performance on the gesture dataset**. We compared our method with the original dense trajectories (DT) on the gesture dataset. "Spatial + DT" indicates "the spatial segmented DT". "Temporal + DT" indicates "the temporal segmented DT". "Spatial + Temporal + DT" indicates "the spatial and temporal segmented DT" i.e., our method.

	Pointing	Nodding	Waving	Mean
Wang et al. [1]	50.0 %	75.0 %	70.0 %	65.0 %
Spatial + DT	65.0 %	90.0 %	75.0 %	76.7 %
Temporal + DT	70.0 %	75.0 %	**80.0 %**	75.0 %
Spatial + Temporal + DT	**80.0 %**	**90.0 %**	75.0 %	**81.7 %**

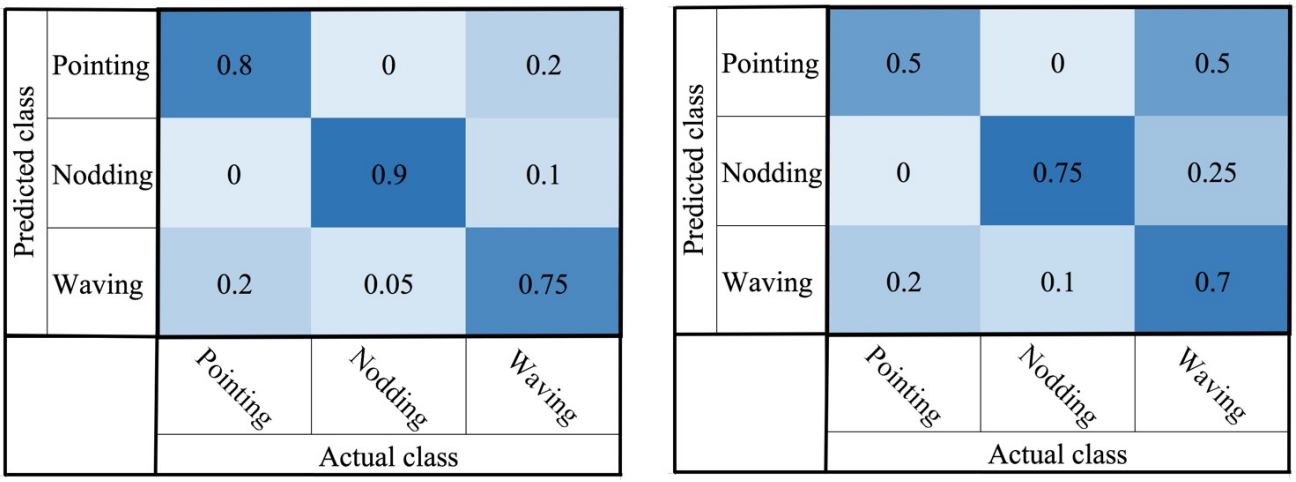
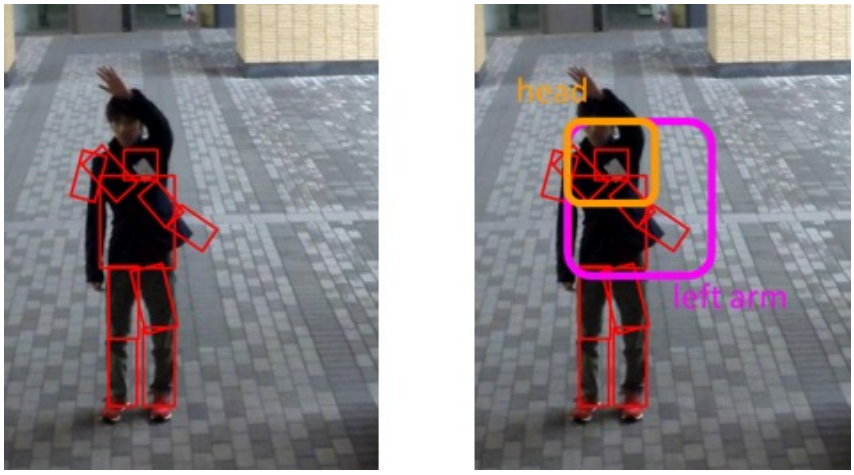


Figure 8. **Confusion matrices for the gesture dataset.** (a): Confusion matrix for the original dense trajectories [1]. (b): Confusion matrix for the proposed method.



(a) (b)

Figure 9. **A failure example of part detection and spatial segmentation.** (a): A failure example of body part detection [2]. (b): A failure example of spatial segmentation.

4. CONCLUSION

This paper improved dense trajectories by extracting local features with spatio-temporal segmentation. To segment spaces, we used a body part detection and improved region range of each body part. To segment times, we divided video frames into fixed time durations. Since only few video datasets are available for gesture classification, we also introduced new gesture dataset. We evaluated our method with the dataset. Experimental results show that our method outperformed the original dense trajectories by about 20% on all of gesture categories. We plan to further improve the performance of gesture recognition by fusing deep-learned features, such as Two-stream Convolutional Networks (ConvNets) which used both RGB ConvNet and optical flow ConvNet for classification.

In future work, we will investigate the performance of group detection by adding gesture recognition as new feature in video. For this purpose, we will also need human pose and eye gaze estimation information in addition to human gesture recognition information.

ACKNOWLEDGEMENTS

This work is supported by the Japan Science and Technology Agency (JST), CREST “Behavior Understanding based on Intention-Gait Model” project.

REFERENCES

[1] Wang, H., Klaser, A., Schmid, C. and Liu, C.-L., "Dense trajectories and motion boundary descriptors for action recognition." IJCV 103.1, 60-79 (2013).
[2] Rothrock, B., Seyoung Park, S., Zhu, S.-C., "Integrating grammar and segmentation for human pose estimation." CVPR, 3214-3221 (2013).
[3] Pellegrini, S., Ess, A. and Van, L.-G., "Improving data association by joint modeling of pedestrian trajectories and groupings." ECCV, 452-465 (2010).

[4] Yamaguchi, K., Berg, A.C., Ortiz, L.E. and Berg, T.L., “Who are you with and where are you going?” CVPR, 1345-1352 (2011).
[5] Ge, W., Collins, R. T., Ruback, R. B., "Vision-based analysis of small groups in pedestrian crowds." TPAMI 34.5, 1003-1016 (2012).
[6] Bazzani, L., Cristani, M. and Murino, V., "Decentralized particle filter for joint individual-group tracking." CVPR, 1886-1893 (2012).
[7] Chamveha, I., Sugano, Y., Sato, Y. and Sugimoto, A., “Social Group Discovery from Surveillance Videos: A Data-Driven Approach with Attention-Based Cues” BMVC (2013).
[8] Scovanner, P., Ali, S. and Shah, M., "A 3-dimensional sift descriptor and its application to action recognition." ACM (2007).
[9] Klaser, A., Marszałek, M. and Schmid, C., "A spatio-temporal descriptor based on 3d-gradients." BMVC, 275-285 (2008).
[10] Willems, G., Tuytelaars, T. and Gool, L., "An efficient dense and scale-invariant spatio-temporal interest point detector." ECCV (2008).
[11] Bregonzio, M., Gong, S. and Xiang, T., "Recognising action as clouds of space-time interest points." CVPR (2009).
[12] Yeffet, L. and Wolf, L., "Local trinary patterns for human action recognition." ICCV (2009).
[13] Matikainen, P., Herbert, M. and Sukthankar, R., “Trajectons: Action recognition through the motion analysis of tracked features.” ICCV Workshops (2009).
[14] Shi, J. and Tomasi, C., "Good Features to Track.", CVPR, 593-600 (1994).
[15] Sun, J., Wu, X., Yan, S., Cheong, L.-F., Chua, T.-S., Li, J., "Hierarchical spatio-temporal context modeling for action recognition." CVPR (2009).
[16] Lowe, D., "Distinctive image features from scale-invariant keypoints." IJCV, 91-110 (2004).
[17] Sun, J., Mu, Y., Yan, S., and Cheong, L.-F. "Activity recognition using dense long-duration trajectories." ICME (2010).
[18] Takahashi, M., Naemura, M., Fujii, M. and Sato, S., "Human action recognition in crowded surveillance video sequences by using features taken from key-point trajectories." CVPR Workshops (2011).
[19] Bay, H., Ess, A., Tuytelaars, T. and Gool, L.V., "SURF: Speeded Up Robust Features." CVIU, 346-359 (2008).
[20] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection." CVPR, 886-893 (2005).
[21] Laptev, I., Marszalek, C., Schmid, C., and Rozenfeld, B., "Learning realistic human actions from movies." CVPR (2008).
[22] Dalal, N., Triggs, B. and Schmid, C., "Human detection using oriented histograms of flow and appearance." European conference on computer vision. Springer Berlin Heidelberg, 2006.
[23] Csurka, G., Dance C.R., Fan, L., Willamowski, J. and Bray, C., "Visual categorization with bags of keypoints." ECCV WORKSHOPS, 1-22 (2004).
[24] Boser, B.E., Guyon, I.M., and Vapnik, V.N., "A training algorithm for optimal margin classifiers." ACM Workshops (1992).
[25] Farnebäck, G. "Two-frame motion estimation based on polynomial expansion." SCIA (2003).
[26] Simonyan, K. and Zisserman, A., "Two-stream convolutional networks for action recognition in videos." NIPS, 568-576 (2014).