Mixture of Heterogeneous Attribute Analyzers for Human Action Detection

Yong Pei, Bingbing Ni(✉), and Indriyati Atmosukarto

Advanced Digital Sciences Center, Singapore, Singapore
{pei.yong,bingbing.ni,indria}@adsc.com.sg

Abstract. We propose a human action detection framework called "mixture of heterogeneous attribute analyzer". This framework integrates heterogeneous attributes learned from static and dynamic, local and global video features, to boost the action detection performance. To this end, we first detect and track multiple people by SVM-HOG detector and tracklet generation. Multiple short human tracklets are then linked into long trajectories by spatio-temporal matching. Human key poses and local dense motion trajectories are then extracted within the tracked human bounding box sequences. Second, we propose a mining method to learn discriminative attributes from these three feature modalities: human bounding box trajectory, key pose and local dense motion trajectories. Finally, the learned discriminative attributes are integrated in a latent structural max-margin learning framework which also explores the spatio-temporal relationship between heterogeneous feature attributes. Experiments on the ChaLearn 2014 human action dataset demonstrate the superior detection performance of the proposed framework.

Keywords: Human trajectory · Key pose · Local dense trajectories · Discriminative mining · Latent structural max-margin learning

1 Introduction

Video-based human action detection has drawn significant research attention in recent years because of its promising applications including smart video surveillance, assisted living, and video-based human computer interaction (HCI). However, human action detection is very difficult given that realistic video sequence contains significant variations of people in posture, motion and clothing, camera motion, view angle changes, illumination changes, occlusions, self-occlusions, and background clutter.

In literature, various types of visual features have been proposed to address this research challenge. Among them, the most popular ones are spatio-temporal motion features, such as spatio-temporal interest points (STIPs) [1] [2] [3] [4] [5] and dense motion trajectories [6] [7]. These features have shown reasonable performance in action recognition. Typically, spatio-temporal motion features are extracted densely over the entire video sequence, and the occurrences of the

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part I, LNCS 8925, pp. 528–540, 2015.
DOI: 10.1007/978-3-319-16178-5_37

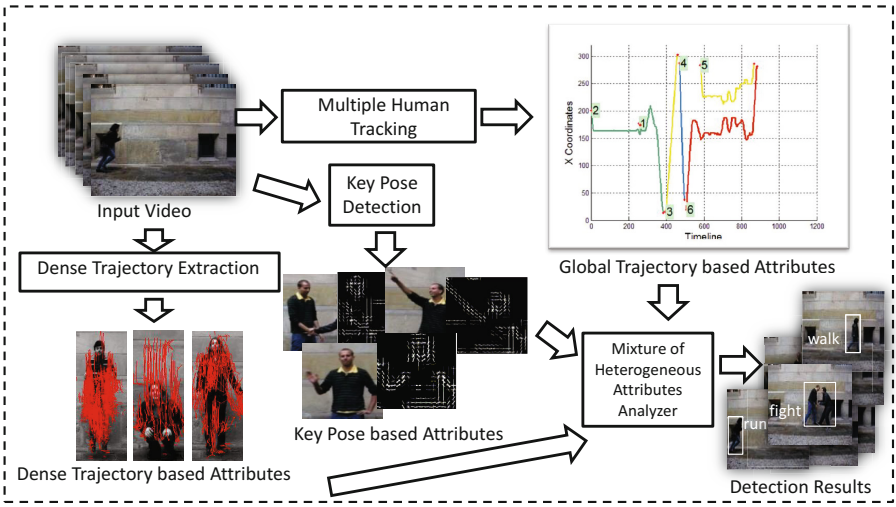

Fig. 1. Overview of the proposed mixture of heterogeneous attribute analyzers scheme for action detection

encoded motion features are accumulated to form the action representation, known as bag-of-words method. Local feature selection are sometimes performed. Wang et al. [8] used a latent structural model for adaptively selecting discriminative local features and their contextual information for action recognition. Ryoo and Aggarwal [9] used a spatio-temporal graph model for selecting and matching $3D$ local interest points and their relationship.

Human key pose information has also been considered for action recognition. Yamato et al. [10] proposed a HMM based method for action recognition by matching frame-wise image features. Lv and Nevatia [11] proposed an action recognition method by key pose matching and Viterbi path searching. In Vahdat et al. [12], sequence of key poses are used for recognizing human interactions. In particular, single human key pose sequences and interactions between two humans key poses are modeled in a graphical model for action recognition. The work by Raptis and Sigal [13] (i.e., which is contemporary with our proposed method) attempts to represent an action with several key frames based on poselet [14] representation.

However, as the problem of action detection is very challenging, previous methods which utilize only one types of visual features might not perform optimally. It is obvious that by combining various types of visual features (heterogeneous features) we can obtain complementary discriminative information and achieve better detection performance. For example, local dense motion trajectories are capable of representing some body part's movement, e.g., hand waving; human key pose is very useful in distinguishing those actions with obvious posture, e.g., two persons in shaking hands; also, human global motion trajectory can tell us whether the person is walking, running or standing still. Motivated by these observations, in this work we propose a novel framework

that integrates heterogeneous attributes learned from static and dynamic, local and global video features, to boost the action detection performance. In particular, we first detect and track multiple people and extract human key poses and local dense motion trajectories features. Second, we propose a mining method to learn discriminative attributes from three feature modalities: human bounding box trajectory, key pose and local dense motion trajectories. Finally, the learned discriminative attributes are integrated in a latent structural max-margin learning framework which also explores the spatio-temporal relationship between heterogeneous feature attributes. Experiments on the ChaLearn 2014 human action dataset: http://gesture.chalearn.org/[15] demonstrate the superior detection performance of the proposed framework.

The rest of this paper is organized as follows. Section 2 presents some related works. Section 3 presents the proposed mixture of heterogeneous attribute analyzers framework for action detection. Extensive experimental results on the ChaLearn 2014 action detection dataset are given in Section 4. Section 5 concludes the paper.

2 Related Work

Previous works integrate human motion and object appearance as well as human interaction information for action recognition. Gupta and Davis [16] proposed a HMM-like Bayesian graphical model to jointly recognize objects and three types of simple movements including *reaching*, *grasping* and *manipulating*. Escorcia and Niebles [17] proposed to represent dynamics of spatio-temporal human interactions using relative object location and size with respect to the human, as well as overlap between human and object, based on pairs of human and object tracks. Prest et al. [18] explored similar idea of modeling the interaction between human and object trajectories, and they proposed a robust human and object tracking method. Different from these works, we proposed to integrate heterogeneous visual feature attributes including local and global, dynamic and static information for action recognition.

3 Methodology

An overview of the proposed mixture of heterogenous feature analyzer based action detection framework is illustrated in Figure 1. Our contributions are as follows. First, we propose a multiple human tracking method and we develop a set of trajectory based visual attributes which can facilitate action recognition. Second, we propose a discriminative key pose mining method to learn informative pose attributes for action representation. Third, we learn dense motion trajectory based attributes to discover discriminative human body part's movement. Finally, we propose a latent structural learning model which integrates heterogeneous visual attributes along with their spatio-temporal relationship for action detection. Details of various components of the proposed framework are elaborated as follows.

3.1 Global Motion Feature: Human Trajectory

We adopt a tracking-by-detection method for tracking human trajectories locally, i.e., to generate short human tracklets. Human bounding boxes are first detected by HOG-SVM detector [19]. Manually labeled human bounding boxes from the training data are used to train the human detector. We use about 3000 training instances and we randomly select a set of negative (non-human) bounding boxes three times of the number of human samples. The bounding box sizes of the training samples are normalized. A scanning window is applied to each video frame to detect the human, and several scales and aspect ratios are used. To improve the human detection accuracy, we iteratively add hard negative samples (i.e., negative samples with high detection scores) detected from the training frames and re-train the model.

We then temporally track the detected human bounding boxes across frames into short segments, i.e., tracklets, based on pairwise matching of human detections over consecutive frames. To do this, we establish all the human detection matches between frame i and $i+1$. To match two detections in consecutive frames, the weighted ℓ_2 distance between their HOG representations and X-Y center coordinates is calculated. We impose that for any detection in frame i, there can be at most one candidate match in frame $i+1$. Those matches which have the length of at least $L_{\min}$ (e.g., 5) frames form human tracklets. To cope with occlusions and missed detections during tracking, we apply an average temporal filter to smooth positions and sizes of the sequence of detected windows, and linear interpolate to fill missed frames.

After we obtain a set of short trajectory segments (tracklets), we spatio-temporally link them into multiple long trajectories for multiple humans in the video. To do this, we first establish the matching between tracklets, using the weighted ℓ_2 distance between their HOG representations and X-Y-T center coordinates between the bounding boxes of the head/tail frames from a pair of tracklets. With n tracklets, we can constructed a $n \times n$ matching graph. We then apply the Hungarian algorithm [20] to obtain the tracked long trajectories.

We then define a set of attributes from the above obtained human bounding box trajectories. Assume each attribute is calculated from a t to $t+T$ temporal window. The attributes defined on a single trajectory include: 1) trajectory length; 2) X-axis moving distance; 3) Y-axis moving distance; 4) mean speed; 5) X-direction mean speed; 6) Y-direction mean speed. The attributes defined on a pair of trajectories include: 1) relative displacement; 2) relative X-axis displacement; 3) relative Y-axis displacement; 4) mean relative speed; 5) mean relative X-direction speed; 6) mean relative Y-direction speed. The final attribute values are obtain by binarizing these values using thresholds: $\phi(x) = \mathrm{I}(\mathrm{x} > \tau)$, where x denotes one of the above defined values and τ is the empirical threshold value estimated from the training data for each type of attribute (i.e., the threshold value is set by maximizing the class separability). All the threshold values can be found from our published code. [1] We denote the vector of the global human trajectory based attributes as $\boldsymbol{\phi}_G$.

[1] Our code will be released upon the publication of this work.

3.2 Static Feature: Human Key Pose

As shown in previous works [11–13], human key pose information can be very helpful in distinguishing different actions, since some actions are associated with some distinctive postures. Inspired by the middle level discriminative mining framework proposed in [21], we develop a discriminative key pose discovery method which contains *seeding*, *re-training* and *selection* phases. This method is described as follows.

Seeding. We obtain the seed key poses using the following pipeline. Training samples are annotated with different aspect ratios and sizes. We first perform a *super-clustering* based on K-means according to the aspect ratios of the annotated samples to divide all training samples into several super clusters. The number of super clusters is typically set as 3. Within each super cluster, we normalize sample size and cluster all training samples into different pose clusters according to their HOG features using K-means. We set the initial number of key pose types K to a large number, i.e., $K = 1000$, since we will select discriminative ones in the later processing step. Each cluster is associated with about 3 to 10 samples. We normalize the average aspect ratio for each cluster and train the linear HOG-SVM model associated with this cluster using the one-*vs*-all scheme. Note that although the same pose type can be shared among multiple action categories, in most cases each cluster only contains instances from the same action class due to the fine granularity being used. The obtained K detection models are regarded as seed key pose types.

Re-training. The purpose of the re-training phase is to consolidate the detection model for each candidate key pose. To do this, we iteratively perform the following steps. We use sliding windows of a key pose detector on the training images to obtain candidate bounding boxes. We then add the top ranked (based on the SVM output score) new instances detected from the images of the related classes to the positive training set and those top ranked new instances from the unrelated classed to the negative training sets (as *hard negative*). We then re-train the HOG-SVM detector. This iteration is performed 10 times (i.e., our empirical study shows 10 is enough).

Selection. Finally we use the entropy which was defined in [21] to select distinctive key pose models. In practice, for each action, we retain the top $K = 3$ to $K = 5$ key pose models.

The key pose attributes are defined as $\boldsymbol{\phi}_P(\mathbf{x}) = [\phi^1(\mathbf{x}), \phi^2(\mathbf{x}), \cdots, \phi^K(\mathbf{x})]^T$, where each $\phi^k(\mathbf{x})$ is a linear detection output. $\mathbf{x}$ denotes the HOG features.

3.3 Local Motion Feature: Dense Trajectory

Dense trajectory has shown its great potential in action recognition [7] [22]. We follow the method in [23] to learn discriminative attributes from the bag-of-words representation. The dictionary size is set as 2000. Max pooling is used. For dense trajectory extraction and descriptor computation, we use the toolbox provided in [7] [22]. Besides the learned dense trajectory attributes, based on

video observations, two types of additional visual attributes are defined for dense trajectories. These attributes include: 1) normalized trajectory length compared to the human bounding box; and 2) principle orientation of trajectory. Details are provided as follows.

Normalized trajectory length. This attribute is useful for describing the magnitude of movement. For example, small movement during the action *Wave Hands* generates relatively short trajectories, while the action *Crouch Down* and *Jump* generate relatively long trajectories. Mathematically, assuming a trajectory is represented by a series of 2D points $\{(x_t, y_t)\}_{i=1,\cdots,T}$ (i.e., T denotes the number of frames), this attribute is defined as $\sum_{i=2}^{T} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} / h_{\text{bbox}}$, where h_{bbox} denotes the height of the detected human bounding box.

Principle orientation of trajectory. Different actions produce different orientated trajectories. For example, actions such as *Wave Hands* and *Clap Hands* mainly produce horizontal orientated trajectories, while *Crouch* and *Jump* will most probably generate vertical ones. The mathematical definition for this attribute is given as: $\sum_{i=2}^{T} |y_i - y_{i-1}| / h_{\text{bbox}}$. Large value means vertical orientated trajectory, and vice versa.

A single trajectory can only provide weak information. Therefore we aggregate the effect of the trajectories in a pooling window, i.e., the number of trajectories with a certain attribute is used for representation. In this work, instead of using a fixed location/scale pooling window, we propose to use a variable pooling window which is shown in Figure 3.

We also design another attribute called **histogram of trajectory points**. As different actions involve different body parts, the densities of trajectories at different positions of the human body varies significantly. Therefore, it is meaningful to use the distribution of dense trajectories for action description. A 3×3 or 5×3 grid on the bounding box of human detection is used to calculate the trajectory point spatial distribution, which is a 9/15-dimensional vector as illustrated in Figure 2. We denote the dense trajectory based attribute vector as ϕ_D.

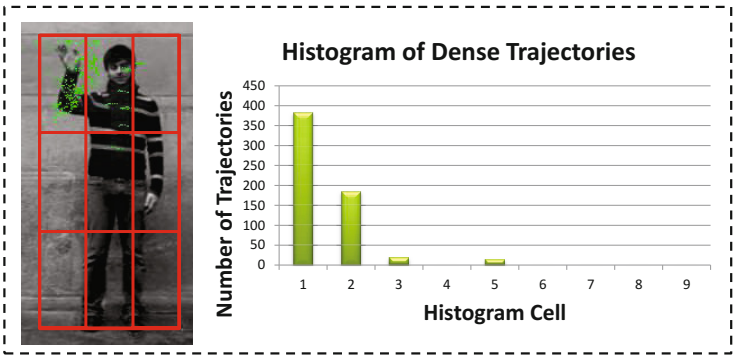

Fig. 2. Illustration of the trajectory point spatial distribution feature

3.4 Heterogeneous Attributes Integration

The above feature extraction step provides us with a set of heterogeneous attributes which contain dynamic and static, local and global visual features. The next step is to seamlessly integrate these attributes for action representation and detection. To this end, we propose a latent structural model for attributes combination. This model not only reflects the discriminative capability of individual attributes, but also encodes the spatio-temporal relationship between different attributes. The model is defined as follows:

$$S(\mathbf{w}, \mathbf{p}, \mathbf{h}) = \mathbf{w}_G^T \boldsymbol{\phi}_G + \mathbf{w}_P^T \boldsymbol{\phi}_P(\mathbf{p}) + \mathbf{w}_D^T \boldsymbol{\phi}_D(\mathbf{h}) + \mathbf{w}_T^T \boldsymbol{\psi}_T(\mathbf{p}) + \mathbf{w}_S^T \boldsymbol{\psi}_S(\mathbf{h}). \quad (1)$$

Each term is explained as follows. $\boldsymbol{\phi}_G$ is the human global trajectory based attribute vector. $\boldsymbol{\phi}_P(\mathbf{p})$ denotes the concatenated HOG key pose representations sampled according to the temporal indices specified by the hidden vector $\mathbf{p}$, i.e., $\mathbf{p} = [t_1; t_2; \cdots; t_K]$ assuming K key poses are selected. $\boldsymbol{\phi}_D(\mathbf{h})$ is the pooled dense motion trajectory based attribute representation and the pooling window is specified by $\mathbf{h}$, i.e., $\mathbf{h} = [dx, dy, dw, dh]$ where (dx, dy) denotes the center offset of the pooling window with reference to the human bounding box and (dw, dh) is the relative width and height of the pooling window with respect to the human bounding box. $\boldsymbol{\psi}_T(\mathbf{p}) = [t_1; t_1^2; t_2; t_2^2; \cdots; t_K; t_K^2]$ measures the temporal configuration of the key poses. $\boldsymbol{\psi}_S(\mathbf{h}) = [dx; dx^2; dy; dy^2; dw; dw^2; dh; dh^2]$ measures the spatial configuration of the dense trajectory attribute pooling window. Please refer to Figure 3 for illustration. The model weights to be learned are defined as $\mathbf{w} = [\mathbf{w}_G; \mathbf{w}_P; \mathbf{w}_D; \mathbf{w}_T; \mathbf{w}_S]$.

The objective function can be rewritten as the linear model $S(\mathbf{w}, \mathbf{p}, \mathbf{h}) = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{p}, \mathbf{h})$, where $\boldsymbol{\varphi}(\mathbf{p}, \mathbf{h}) = [\boldsymbol{\phi}_G; \boldsymbol{\phi}_P(\mathbf{p}); \boldsymbol{\phi}_D(\mathbf{h}); \boldsymbol{\psi}_T(\mathbf{p}); \boldsymbol{\psi}_S(\mathbf{h})]$. Assume we have N labeled training samples and y^t the corresponding action label for sample t. The model learning problem is formulated as:

$$\begin{aligned} &\min_{\mathbf{w}, \boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + \mathcal{C} \sum_{t=1}^{N} \xi_t, \\ &s.t. \ \max_{\mathbf{h}, \mathbf{p}} \ \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{p}, \mathbf{h}) \geq 1 - \xi_t, \text{if } \mathrm{y^t} = 1, \xi_\mathrm{t} \geq 0, \forall \mathrm{t}, \\ &\qquad \mathbf{w}^T \boldsymbol{\varphi}(\hat{\mathbf{p}}, \hat{\mathbf{h}}) \leq -1 + \xi_t, \ \forall \hat{\mathbf{h}}, \hat{\mathbf{p}}, \text{if } \mathrm{y^t} = -1, \xi_\mathrm{t} \geq 0, \forall \mathrm{t}. \end{aligned} \quad (2)$$

Here $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_N)^T$ denotes the set of slack variables and $\mathcal{C} = 1000$ the weighting factor for the constraints. This latent structural SVM model leads to a non-convex optimization problem. We follow the cutting plane based optimization scheme proposed in [24] for model learning, which has been widely used in latent structural learning problems [25] [8].

For action detection, we run temporal sliding windows with variable length (i.e., fixed set of lengths) and optimize the configuration $\mathbf{h}$ and $\mathbf{p}$ by dynamic programming. For detection efficiency, we use a fixed set of spatial pooling window configurations.

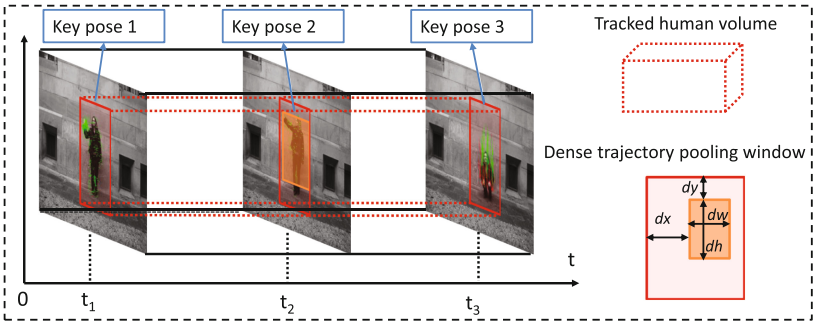

Fig. 3. Illustration of the spatio-temporal relationship encoding used in our proposed latent structural learning formulation. Note that dx, dy, dw, dh are scaled with respect to the human detection bounding box (red).

4 Experiment

In this section, we will provide systematic evaluations on the effectiveness of our proposed attribute learning modules, as well as the entire framework for action detection.

4.1 Dataset

The dataset used for evaluation is the ChaLearn 2014 Track 2 action recognition dataset, which focuses on action/interaction recognition on RGB data. The dataset contains a labeled database of 235 action performances from 17 users. It includes 11 action categories: *Wave*, *Point*, *Clap*, *Crouch*, *Jump*, *Walk*, *Run*, *Shake Hands*, *Hug*, *Kiss*, and *Fight*. Seven sequences are used for training (five training sequences and two validation sequences) and two sequences are used for testing. The training set contains 150 manually labeled action performances as well as a validation dataset with 90 labeled action performances. The final evaluation data (testing set) contains 95 performances.

The evaluation metric used for the Chalearn 2014 dataset is based on the Jaccard Index. Readers who are interested in the details of the metric please refer to the ChaLearn 2014 data description website: http://gesture.chalearn.org/mmdata.

4.2 Heterogenous Attribute Extraction Results

We first show the results of feature extraction for various types (heterogeneous) of visual features including global human trajectory, key pose and dense motion trajectory. Figure 4 visualizes several examples of the tracked multiple human global moving trajectories from the ChaLearn 2014 dataset. Left images (w.r.t the dash line) show the tracked short trajectories and right ones show the linked long trajectories. We annotate different tracklets with different numbers. From

Figure 4 we note that although the proposed multiple human tracking module is very simple, the tracking results are quite reasonable. Also, the proposed tracklet linking method is quite effective and it can remove spurious tracklets. We will see in the rest of the section that the global trajectory based attributes obtained using this proposed tracking method greatly help the action recognition task.

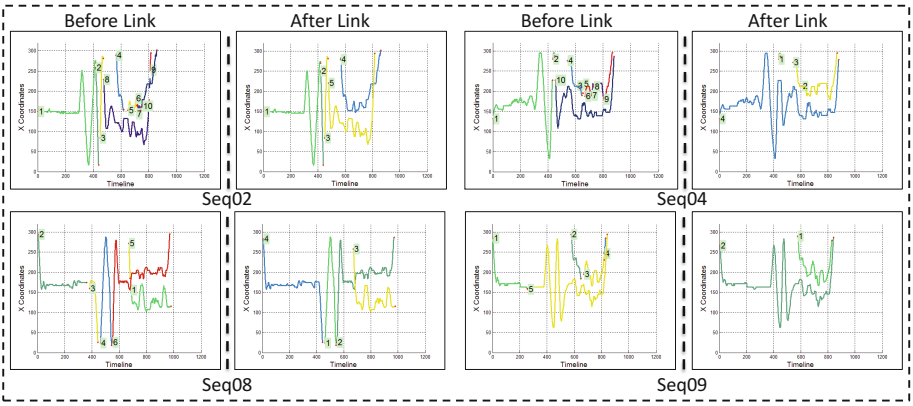

Fig. 4. Examples of the multiple people tracking results in terms of the X direction movement. Images on the left of the dash line show the short tracklets using our human detection and tracking method; images on the right of the dash line show the linked long trajectories of multiple persons using our tracklet matching method.

In Figure 5, we illustrate several examples of the learned discriminative key poses with high entropy values (each column corresponds to the instances belonging to one key pose) using our proposed mining algorithm. We see that these mined key poses are quite consistent within each cluster and they are representative poses for certain actions. For example, the key pose 1, 2, 3,4 well correspond to the action *Kiss/Hug*, *Point*, *Crouch* and *Shake Hands*, respectively. Note that a learned key pose can be shared by more than one action category. We will show in the later experiments that these learned discriminative key poses play an important role in action recognition.

In Figure 6, we show the temporal distributions of several example attributes based on dense motion trajectories on Sequences 04 of the ChaLearn 2014 dataset. X-axis represents the time line and the Y-axis value denotes the number of dense trajectories that possess the concerned attribute. Note that the peaks of attribute 1 and 2 in the left figure correspond to the action *Crouch* and *Sit up* respectively, while the peak of attribute 3 in the right figure corresponds to the action *Jump*. This result shows that extracted dense trajectory based attributes reflect discriminative information.

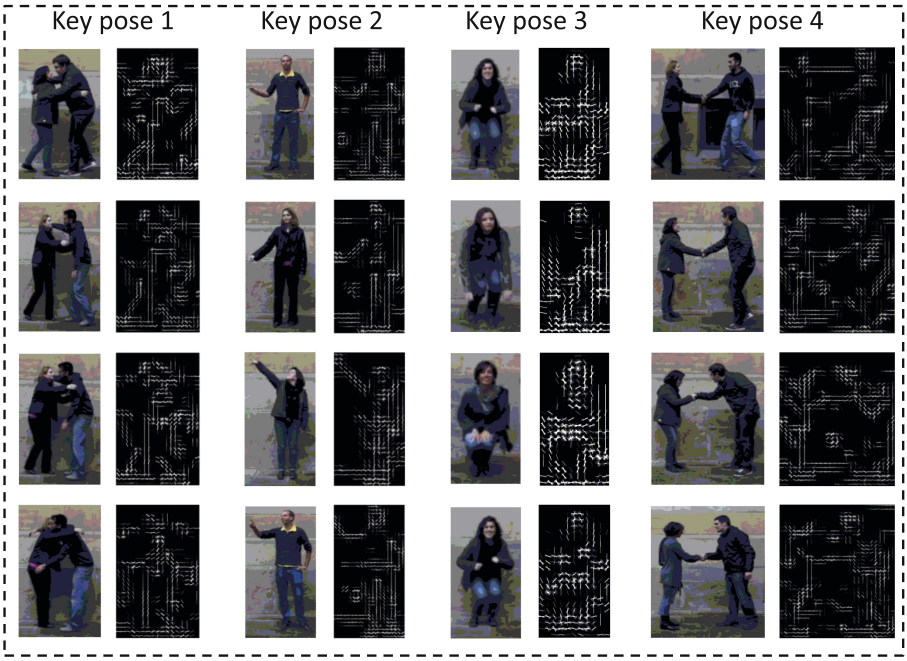

Fig. 5. Examples of the learned discriminative key poses

4.3 Action Detection Results

We compare our method with the state-of-the-art action recognition method, i.e., the dense trajectory method [7] [22]. We use the implementation and the default parameter settings provided by the authors [7] [22]. We do not compare with the recently proposed improved dense trajectory method [26] since on the ChaLearn 2014 dataset, the background motion is ignorable therefore the original dense trajectory method can perform equally well as the improved dense trajectory method. The action recognition comparison result is shown in Figure 7. We note from Figure 7 that the proposed mixture of heterogeneous analyzers method greatly outperforms the dense trajectory based method. This is because that the proposed method seamlessly combines heterogeneous features such as global motion, local dense trajectory and static key pose feature, which explores the complementary nature among all these types of visual features and boosts the recognition performance. On the contrary, using dense trajectory only is not optimal in dealing with some action classes. As can be seen from Figure 7, key pose information is a more discriminative cue for recognizing the action *Hug* and *Shake Hands* as these two classes do not contain rich dense trajectory features. In general, combining various visual feature attributes achieves the best action detection performance. Our final evaluation score on the testing data of ChaLearn 2014 action dataset is 0.501164.

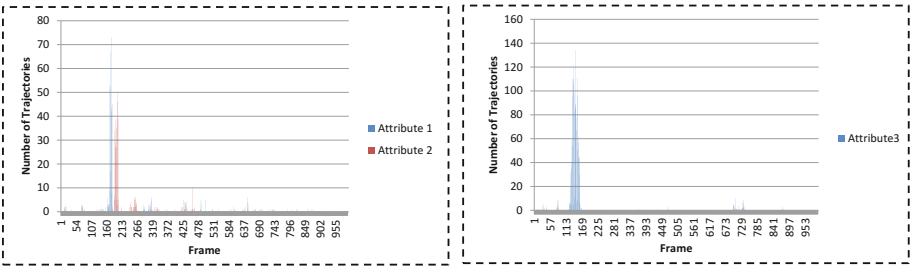

Fig. 6. Temporal distribution (histogram) for three example attributes based on the dense trajectory features on Sequence 04 of the ChaLearn 2014 action dataset. The peaks of distributions correspond to the action *Crouch*, *Sit up* and *Jump*, respectively.

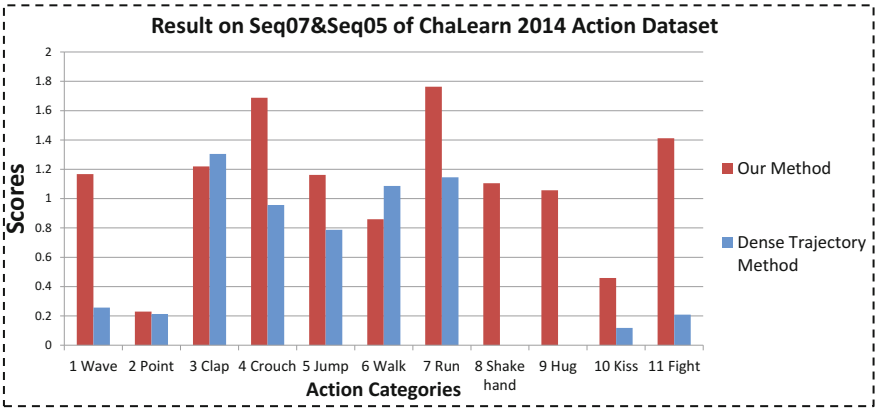

Fig. 7. Per class action recognition performance comparison on the ChaLearn 2014 dataset (testing set). Our final evaluation score on the testing data of ChaLearn 2014 action dataset is 0.501164.

To further unveil the working mechanism of the proposed mixture of heterogeneous analyzers scheme, we conduct a study to evaluate the effectiveness of different components (attributes) of our method including: global human trajectory based attribute, key pose based attribute and dense trajectory based attribute. Namely, for detecting some actions, we disable some types of attributes and evaluate the detection performance drop. The comparison results are shown in Table 1. From Table 1, we note that in general, combining different types of visual attributes achieves much better action recognition performance. For some actions such as *Hug*, key pose based attributes possess discriminative capabilities since almost every two people have the similar way of performing *Hug*. For actions such as *Wave*, dense motion trajectories play an important role in classification. This is because different persons have different ways of pointing with different poses and the more important cue is the local movement induced by hand motion.

Table 1. Action detection performances using different combinations of heterogeneous attributes

Sequence No.	Action	Method	Recall	Accuracy	Score
07	Hug	global trajectory	0.6452	0.2985	0.2564
		global trajectory + key pose	0.6452	1.0000	0.6452
05	Wave	key pose	0.3600	1.0000	0.3600
		key pose + dense trajectory	1.0000	0.8333	0.8333
05	Point	key pose	0.8333	0.1005	0.0985
		key pose + global trajectory	0.8333	0.2500	0.2381

5 Conclusion

We have proposed a mixture of heterogenous feature analyzers scheme that integrates various types of visual features including static and dynamic, local and global features and explores their spatio-temporal relationship, for discriminative action representation. Extensive experiment on the ChaLearn 2014 action datasets demonstrates the effectiveness of the proposed heterogeneous feature integration framework.

Acknowledgments. The study is supported by a research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research(A*STAR).

References

1. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (2005)
2. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d gradients. In: British Machine Vision Conference (2008)
3. Laptev, I., Lindeberg, T.: Space-time interest points. In: International Conference on Computer Vision (2003)
4. Niebles, J.C., Chen, C.-W., Fei-Fei, L.: Modeling temporal structure of decomposable motion segments for activity classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 392–405. Springer, Heidelberg (2010)
5. Tang, K., Fei-Fei, L., Koller, D.: Learning latent temporal structure for complex event detection. In: International Conference on Computer Vision and Pattern Recognition (2012)
6. Raptis, M., Kokkinos, I., Soatto, S.: Discovering discriminative action parts from mid-level video representations. In: International Conference on Computer Vision and Pattern Recognition (2012)
7. Wang, H., Kläser, A., Schmid, C., Cheng-Lin, L.: Action recognition by dense trajectories. In: International Conference on Computer Vision and Pattern Recognition, pp. 3169–3176 (2011)
8. Wang, Y., Mori, G.: Hidden part models for human action recognition: Probabilistic versus max margin. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(7), 1310–1323 (2011)

9. Ryoo, M.S., Aggarwal, J.: Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In: International Conference on Computer Vision, pp. 1593–1600 (2009)
10. Yamato, J., Ohya, J., Ishii, K.: Recognizing human action in time-sequential images using hidden markov model. In: International Conference on Computer Vision and Pattern Recognition, pp. 379–385 (1992)
11. Lv, F., Nevatia, R.: Single view human action recognition using key pose matching and viterbi path searching. In: International Conference Computer Vision and Pattern Recognition (2007)
12. Vahdat, A., Gao, B., Ranjbar, M., Mori, G.: A discriminative key pose sequence model for recognizing human interactions. In: ICCV Workshop, pp. 1729–1736 (2011)
13. Raptis, M., Sigal, L.: Poselet key-framing: A model for human activity recognition. In: International Conference on Computer Vision and Pattern Recognition, pp. 2650–2657 (2013)
14. Bourdev, L., Malik, J.: Poselets: body part detectors trained using 3d human pose annotations. In: International Conference on Computer Vision (2009)
15. Snchez, D., Bautista, M., Escalera, S.: Hupba 8k+: Dataset and ecoc-graphcut based segmentation of human limbs. Neurocomputing (2014)
16. Gupta, A., Davis, L.: Objects in action: an approach for combining action understanding and object perception. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
17. Escorcia, V., Niebles, J.: Spatio-temporal human-object interactions for action recognition in videos. In: IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 508–514 (2013)
18. Prest, A., Ferrari, V., Schmid, C.: Explicit modeling of human-object interactions in realistic videos. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(4), 835–848 (2013)
19. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: International Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
20. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly **2**, 83–97 (1955)
21. Juneja, M., Vedaldi, A., Jawahar, C.V., Zisserman, A.: Blocks that shout: distinctive parts for scene classification. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
22. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision **103**(1), 60–79 (2013)
23. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3337–3344 (2011)
24. Joachims, T., Finley, T., Yu, C.N.J.: Cutting-plane training of structural svms. Machine Learning **77**(1), 27–59 (2009)
25. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9), 1627–1645 (2010)
26. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: International Conference on Computer Vision (2013)