Action Recognition Robust to Background Clutter by Using Stereo Vision

Jordi Sanchez-Riera, Jan Čech, and Radu Horaud

INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, FRANCE
{jordi.sanchez-riera, jan.cech, radu.horaud}@inria.fr

Abstract. An action recognition algorithm which works with binocular videos is presented. The proposed method uses standard bag-of-words approach, where each action clip is represented as a histogram of visual words. However, instead of using classical monocular HoG/HoF features, we construct features from the scene-flow computed by a matching algorithm on the sequence of stereo images. The resulting algorithm has a comparable or slightly better recognition accuracy than standard monocular solution in controlled setup with a single actor present in the scene. However, we show its significantly improved performance in case of strong background clutter due to other people freely moving behind the actor.

1 Introduction

An extensive research has been done in action recognition throughout recent years, which is well documented in survey papers [1, 2]. Most of the methods work with monocular videos only. Very successful methods use image retrieval techniques, where each video sequence is represented as a histogram of visual words [3], and large margin classifier is then used for recognition.

In particular, spatiotemporal interest points [3] are detected in the image sequence. These points are described by a descriptor HoG (Histogram of Gradients)/HoF (Histogram of Optical Flow) [4] which captures the surrounding of an interest point. The descriptors are quantized by K-means clustering and each videoclip is represented as a histogram with K bins. Support Vector Machine is then used for classification.

Further research to improve the recognition accuracy went in the direction of densifying the interest points and enhancing the local descriptors. The interest points employed in [3] are spatiotemporal extensions of a Harris corner detector, i.e. locations in a video stream having large local variance in both spatial and temporal dimensions, representing abrupt events in the stream. This is in order to achieve high repeatability of the detection. However, such points are quite rare and important relevant information can be missed. Therefore there were alternatives to these interest points, e.g. based on Gabor filters [5, 6], or even simply using a regular dense sampling [7] to reach higher coverage, or a hybrid scheme by [8], which start by dense sampling and optimize the position and scale within a bounded area in order to increase the coverage and preserve the

repeatability of the interest points. An extension of the original HoG/HoF descriptor was proposed e.g. by spatiotemporal gradients [9], or motion boundary histograms [10].

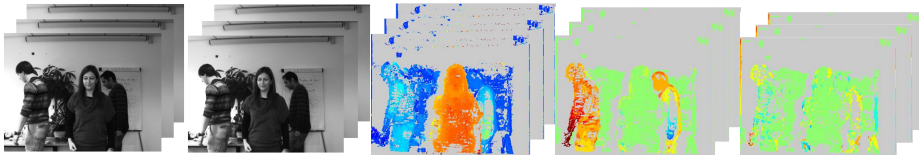
However these methods can be quite sensitive to background clutter present in populated scenes, since interest points are detected not only in the actor but on the background as well. This causes the global histogram representation to be corrupted and the accuracy is significantly decreased.

Stereo vision or multiple view vision have not been much used in action recognition. Using stereo, the existing methods typically try to make the algorithm insensitive to a camera viewpoint [11]. Similarly [12] uses a special room and a multi-camera setup to construct viewpoint invariant action representation, and [13] incorporate temporal information to the multi-view setup. Work [14] uses the depth map obtained by stereo matching to fit an articulated body model and use joint trajectories for action recognition.

An alternative to stereo vision is using RGB-D sensor, which provides a depth image besides the color/intensity image. It is based on time-of-flight or structured light technology. This research is vivid nowadays due to the recent irruption of Kinect device. For instance [15] constructs 3D motion primitives from a cloud of 3D points. Work [16] extends 2D silhoutte by projection of the point cloud into three orthogonal planes. In [17] the authors uses local interest point descriptors which are computed from spatiotemporal image and depth gradients for each pixel of a spatiotemporal neighbourhood of interest points. Since the neighbourhood is large, they use PCA for dimensionality reduction prior to quantization. In [18], spatiotemporal interest points are divided into different layers based on depth and a multichannel histogram is created. Another direction is to estimate the body skeleton from the depth data. Commercially successful real-time game controller uses skeleton model from body part labelling of depth data of Kinect [19]. Joint trajectories are used for action or gesture recognition in e.g. [20, 21]. However, for some applications such active sensors are not suitable. For example, in outdoor setup or in a scenario with multiple autonomous robots whose active sensors would interfere to each other.

Therefore we propose a simple stereo vision based method, which can focus the algorithm to an active actor while disregarding the background activity based on completely passive system, see Fig. 1. Our contribution is extending the original successful action recognition framework [3] with descriptors based on stereo-vision and the scene-flow. We observed a significant improvement of the proposed method in the robustness to the perturbations due to the uncontrolled motion of other people behind the actor.

The rest of the paper is structured as follows: The proposed method is described in detail in Sec. 2. Experimental validation which includes a comparison with the state-of-the-art algorithm is presented in Sec. 3. Finally, Sec. 4 concludes the paper.



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

Fig. 1: Example of data for one sequence. The input data consists of sequences of (a) Left and (b) Right images. The maps of (c) disparity, (d) horizontal, (e) vertical component of the optical flow computed by algorithm [22]. The maps are color-coded: gray color means unassigned value, for disparity warmer colors corresponds to points closer to the camera, for optical flow warmer colors corresponds to motion to the left and up respectively.

2 Method Description

Before we give details on the proposed descriptor, we briefly revise the bag-of-words (BoW) paradigm for action recognition. Following [3] it requires to:

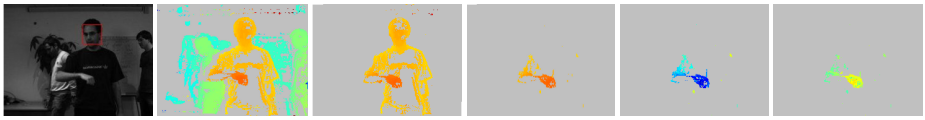
1. Collect a set of local descriptors associated to the interest points for image frames of all training action video clips.
2. Apply clustering algorithm to these descriptors, for instance, K-means.
3. Quantize the descriptor to get the 'visual words'. For each descriptor, assign label according to its nearest cluster centroid.
4. Represent a video clip as a K-bins histogram of the quantized descriptor ('bag of words').
5. Train a classifier with these histograms, for instance, SVM.

In Steps 1–3, the the visual word vocabulary (or the codebook) is constructed. The dimensionality of the local descriptor is typically high and the space is consequently sparse, that is why it is represented by K clusters of observed data. In Steps 4–5, a compact (K-length vector) representation of training videoclips with annotated labels is used to train a classifier. The 'bag of words' representation encodes a relative frequence of occurences of the quantized desriptors and it turns out to be discriminative among action classes. Later for recognition, an unknown videoclip is first represented as the K-length histogram and then it is fed to the classifier which assigns the class label.

We follow exactly this framework, except for the Step 1. Unlike the monocular HoG/HoF descriptor [3], we introduce a new descriptor based on the Scene-Flow [22].

2.1 Local Descriptor based on the Scene-flow

The Scene-flow is a 3D extension of the optical flow. We represent a scene-flow as depth and optical flow, which together with a camera calibration is equivalent to



(a) (b) (c) (d) (e) (f)

Fig. 2: Construction of the proposed descriptor. The actor's face is detected from the left input image (a). The raw disparity map (b) is segmented, such that all pixels having the lower disparity than the actor's face are discarded (c). The descriptor is then computed for all remaining pixels undergoing non-zero motion, such that it consists of the pixel's position relative to the face, it's disparity (d), and horizontal (e) and vertical (f) components of optical flow.

a vector field of 3D position and associated 3D velocities of reconstructed surface points. This intrinsic representation is potentially less sensitive to the changes of texture and illumination in the action dataset than the representation which relies solely on the intensity images. Moreover, with the notion of depth, it is straightforward to focus the actor performing the action to be recognized while discarding any activity from the background clutter.

We assume the action performing actor is the person which is the closest to the camera. We believe this is a reasonable assumption, which is typically the case of human-robot interaction or movies.

The proposed descriptor is constructed as follows, see Fig. 2:

1. Get the synchonized sequences of the left $\mathbf{I}_l$ and right images $\mathbf{I}_r$. For each frame compute the disparity map $\mathbf{D}$ and optical flow maps $\mathbf{F}_h, \mathbf{F}_v$ by the algorithm [22].
2. Find the actor's face with a face detector [23]: $(x_0, y_0) = \mathrm{FD}(\mathbf{I}_l)$. In case of multiple faces detected, the one with the highest disparity $d_0 = \mathbf{D}(x_0, y_0)$ is selected[1]. In case no face is detected, if the actor turns or the detector miss the face, we simply assume a previous face position.
3. Segment the scene using disparity and optical flow: (1) Only pixels with magnitude of optical flow greater than zero are considered, (2) Only pixels with disparity greater or equal to the disparity of the actor's face are considered. So the set of valid pixels

$$S = \{(x, y) : \mathbf{F}_h(x, y)^2 + \mathbf{F}_v(x, y)^2 > 0 \text{ and } \mathbf{D}(x, y) > d_0 - \mu\},$$

where $\mu = 5$ is a small margin to ensure the entire actor's body is included.
4. At each reconstructed pixel passing the above test $(x, y) \in S$, the local descriptor is 5-dimensional only:

$$L(x, y) = \Big(x - x_0, y - y_0, \mathbf{D}(x, y) - d_0, \mathbf{F}_h(x, y), \mathbf{F}_v(x, y)\Big).$$

[1] The disparity of the face is estimated as an average disparity inside the bounding box obtained from the face detection. The center of the bounding box is the pixel (x_0, y_0).

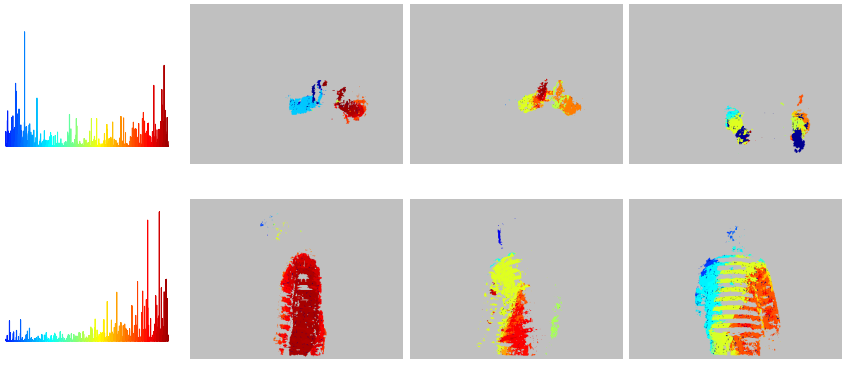


Fig. 3: Histograms of visual words and corresponding assignment to pixels for frames of two actions: clap (top) and turn-around (bottom). The color encodes the indices of visual words $1, \ldots, K$. The coloring is such that similar visual words have similar color. We can see typical visual words occuring during the actions.

Notice the face-normalized position of the pixels, brings a kind of global information into the local descriptor.

Following the BoW procedure described above, after building the codebook and subsequent quantization of pixel descriptors, the resulting histograms of their occurrences in the action video sequence intuitively encodes the activity of actor's body parts in the sense of 3D motion. See Fig. 3 for an illustration.

3 Experiments

To evaluate the performance of the proposed binocular method and compare it with a state-of-the-art monocular method [3], we use the Ravel[2] dataset [24]. The Ravel dataset consists of 7 actions (talk phone, drink, scratch head, turn around, check watch, clap, cross arms) performed by 12 actors in 6 trials each. First 3 trials are with stable static background without other people in the scene (we denote as 'Controlled'), while next 3 trials are performed with motion background clutter due to arbitrary activity of the people behind the actor (we denote as 'Clutter'). See Fig. 4 and Fig. 5 for respective examples. The dataset is challenging due to the strong intra-class variance, strong dynamic background in the 'Clutter', and unstable lighting conditions.

We will show results of two baseline algorithms. The first one is the algorithm described in [3] works with monocular (left camera) stream only and uses the sparse spatiotemporal interests points and HoG/HoF descriptors, we denote as 'STIPs'. The other baseline is the same algorithm, however we ran it in both left and right camera sequences, matched the detected points along the epipolar lines, and removed the interest points which have smaller disparity than the

[2] http://ravel.humavips.eu

Fig. 4: Ravel dataset examples - controlled setup. Note that different actors perform the same action quite differently as for example in "cross arms". Actions: "cross arms", "check watch", "scratch head", "cross arms", "talk phone", "cross arms", "scratch head", "clap".

disparity of the actor's face. The motivation behind is to remove the irrelevant interest points detected on the background clutter. The rest of the algorithm [3] remains the same. We call this algorithm 'STIPs-stereo'. The proposed method described in Sec. 2, is denoted as '5DF'.

The codebook was built in a sequence of a single actor, namely 'character-09'. This actor was not later used either for learning a classifier or for testing. We believe a single actor performing the same set of actions as all other actors sweeps the space of local descriptors is enough and also K-means algorithm is run only once and not in the leave-one-out loop (see later), which would be too time consuming. The size of the codebook K was optimized for all the methods in the logarithmic range from $K = 10$ to $K = 10000$ and the optimum was found for $K = 1000$, the same for all the methods.

Learning a classifier and testing was performed in a standard leave-one-actor-out scenario. One actor was removed from the set, the linear SVM classifier was trained in the sequences of remaining actors and then tested on the sequence of the left actor and this was repeated for all actors. The recognition rate reported is the average error over all actors.

Results are shown in Tab. 1. We can see the proposed method (5DF) performs comparably in the setup when there is a single actor in the scene only. This proves the proposed descriptor computed in the meaningful semi-dense locations is informative. Furthemore, we can see the recognition accuracy of the proposed method does not drop much in cases of the background clutter of other people freely moving behind the actor. This demonstrates that the algorithm can properly focus the active actor while disregarding the background activity using the depth information from stereo. The monocular baseline method [3] (STIPs) is naturally very sensitive to this type of the background clutter. The algorithm cannot distinguish the informative interest points of the clutter from

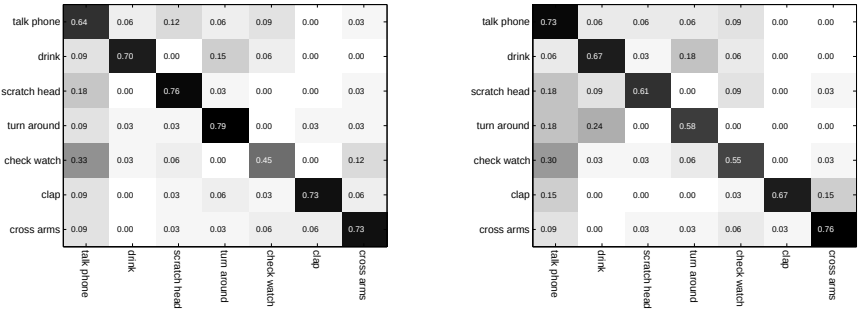
Fig. 5: Ravel dataset examples - cluttered setup. Actions: "turn around", "clap", "talk phone", "talk phone", "turn around", "drink", "check watch", "drink". Note different illumination conditions.

Algorithm	Controlled	Clutter
STIPs [3]	0.6883	0.4675
STIPs-stereo	0.6537	0.5238
5DF (the proposed method)	0.6840	0.6494

Table 1: Recogntion accuracy of the tested methods. The proposed (5DF) method has comparable results with state-of-the-art method (STIPs) in the controlled setup with only one actor in the scene, while it much less sensitive to the strong dynamic background clutter. The other baseline (STIPs-stereo) is less sensitive to the background by using the stereo information, however due to insufficient coverage of interest points the recognition accuracy is lower.

corresponding descriptors on other people in the scene, which contaminates the histograms and the recognition accuracy drops significantly. The second baseline (STIPs-stereo), which attempts to remove the interest points detected on the background by stereo matching, is less sensitive to the background clutter, however its recognition accuracy is slightly lower for 'controlled' setup. The reason is that the sparse spatiotemporal interest points become even sparser, since the stereo matching may discard also points on the foreground due to matching ambiguity. Notice that in STIPs method, we have about 10 interest points per frame, but in our method we have about 10000 locations per frame where descriptors are computed.

For more insight, we show confusion matrices of both methods for both 'controlled' and 'clutter' setups, see Fig. 6–8. For instance, we can see that scratch head is confusing with talk phone. This is not so surprising since these actions starts with the hand at the level of the pocket and is directed to the head, where



(a) Controlled (b) Clutter

Fig. 6: Confusion Matrix for the proposed method (5DF) for a) Controlled and b) Cluttered setup.

the difference is whereas it remains static (talk phone) or moving (scratch head). Again, there is significantly much less confusion in case of the background clutter in the proposed binocular method compared to the state-of-the-art method which only uses a monocular video. This corroborates that stereo vision brings an important extra information.

4 Conclusion

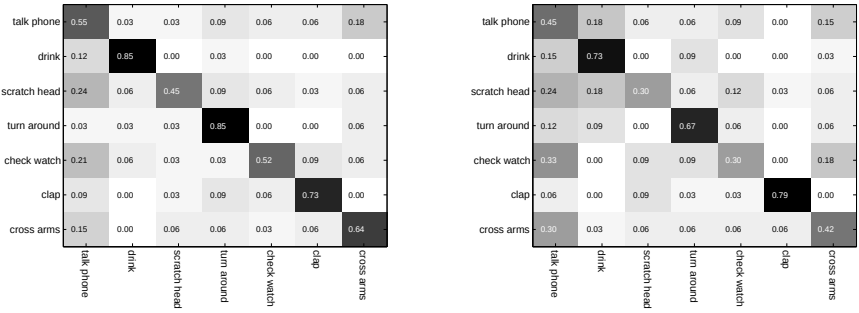
We presented an action recognition method which uses the scene-flow computed from binocular video sequences. Experimentally we proved that the extra information from stereo significantly improves the recognition accuracy in the presence of strong background clutter.

The proposed method requires the actor's face is detected in majority of the frames. We expect that a tracker with a motion model would help to localize the face if it is turned away. Future work includes an elaboration on the design of the local descriptor. Combination of the local descriptor with the proposed one could further improve the recognition accuracy.

Acknowledgements. This research was supported by EC project FP7-ICT-247525-HUMAVIPS.

References

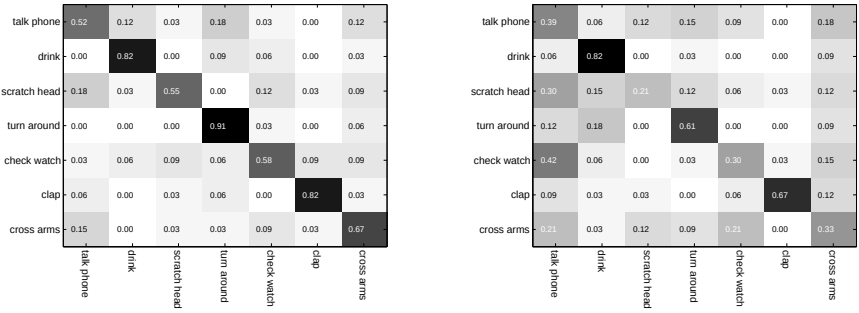
1. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. CVIU **115** (2011) 224–241
2. Poppe, R.: A survey on vision-based human action recognition. IVC **28** (2010) 976–990



(a) Controlled (b) Clutter

Fig. 7: Confusion Matrix for the STIPs-stereo for a) Controlled and b) Cluttered setup.

3. Laptev, I.: On space-time interest points. IJCV **64** (2005)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. CVPR. (2005)
5. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. VS-PETS (2005)
6. Bregonzio, M., Gong, S., Xiang, T.: Recognising action as clouds of space-time interest points. In: Proc. CVPR. (2009)
7. Wang, H., Klaser, A., Laptev, I., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: Proc. BMVC. (2009)
8. Tuytelaars, T.: Dense interest points. In: Proc. CVPR. (2010)
9. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: Proc. BMVC. (2008)
10. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Proc. CVPR. (2011)
11. Roh, M.C., Shin, H.K., Lee, S.W.: View-independent human action recognition with volume motion template on single stereo camera. Pattern Recognition Letters **31** (2010) 639–647
12. Weinland, D., Boyer, E., Ronfard, R.: Action recognition from arbitrary views using 3D exemplars. In: Proc. ICCV. (2007)
13. Yan, P., Khan, S.M., Shah, M.: Learning 4D action feautre models for arbitrary view action recognition. In: Proc. CVPR. (2008)
14. Uddin, M.Z., Thang, N.D., Kim, J.T., Kim, T.S.: Human activity recognition using body joint-angle features and hidden Markov model. ETRI Journal **33** (2011) 569–579
15. Holte, M.B., Moeslund, T.B., Fihl, P.: View-invariant gesture recognition using 3d optical flow and harmonic motion context. CVIU **114** (2010) 1353–1361
16. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: Proc. CVPR workshop on Human Communicative Behaviour Analysis. (2010)
17. Zhang, H., Parker, L.E.: 4-dimensional local spatio-temporal features for human activity recognition. In: Proc. IROS. (2011)



(a) Controlled (b) Clutter

Fig. 8: Confusion Matrix for the state-of-the-art method [3] (STIPs) for a) Controlled and b) Cluttered setup.

18. Ni, B., Wang, G., Moulin, P.: RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In: Proc. ICCV Workshop on Consumer Depth Cameras for Computer Vision. (2011)
19. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M.: Real-time human pose recognition in parts from single depth images. In: Proc. CVPR. (2011)
20. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from rgbd images. In: Proc. ICRA. (2012)
21. Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3D joints. In: Proc. CVPR workshop on Human Activity Understanding from 3D Data (HAU3D). (2012)
22. Cech, J., Sanchez-Riera, J., Horaud, R.P.: Scene flow estimation by growing correspondence seeds. In: Proc. CVPR. (2011)
23. Šochman, J., Matas, J.: Waldboost – learning for time constrained sequential detection. In: CVPR. (2005)
24. Alameda-Pineda, X., Sanchez-Riera, J., Franc, V., Wienke, J., Cech, J., Kulkarni, K., Deleforge, A., Horaud, R.P.: Ravel: An annotated corpus for training robots with audiovisual abilities. In: Journal on Multimodal User Interfaces. (2012)