January 2014

ACTION RECOGNITION USING FEATURE TRANSFORM DESCRIPTOR FROM MINED DENSE SPATIO TEMPORAL

N. R . VIKRAM

Department of information technology, M.I.T campus, Anna University, Chennai, nrvikram89@gmail.com

P. M. ASHOKKUMAR

Department of information technology, M.I.T campus, Anna University, Chennai, pmashokk@gmail.com

Follow this and additional works at: <https://www.interscience.in/ijcsi>

Part of the [Computer Engineering Commons](#), [Information Security Commons](#), and the [Systems and Communications Commons](#)

Recommended Citation

VIKRAM, N. R . and ASHOKKUMAR, P. M. (2014) "ACTION RECOGNITION USING FEATURE TRANSFORM DESCRIPTOR FROM MINED DENSE SPATIO TEMPORAL," *International Journal of Computer Science and Informatics*: Vol. 3 : Iss. 3 , Article 13.

DOI: 10.47893/IJCSI.2014.1148

Available at: <https://www.interscience.in/ijcsi/vol3/iss3/13>

This Article is brought to you for free and open access by the Interscience Journals at Interscience Research Network. It has been accepted for inclusion in International Journal of Computer Science and Informatics by an authorized editor of Interscience Research Network. For more information, please contact sritampatnaik@gmail.com.

ACTION RECOGNITION USING FEATURE TRANSFORM DESCRIPTOR FROM MINED DENSE SPATIO TEMPORAL

N. R. VIKRAM¹ & P. M. ASHOKKUMAR²

^{1,2}Department of information technology, M.I.T campus, Anna University, Chennai
E-mail : ¹nrvikram89@gmail.com, ²pmashokk@gmail.com

Abstract - Action recognition in video sequence has been a major challenging research area for number of years. Apriori algorithm and SIFT descriptor based approach for action recognition is proposed in this paper. Here, two phases can be carried out for accurate and updating of action recognition. In the first phase, the input should be the video sequence. For preprocessing the sequences frame can be formatted by background modeling for every successive frame. After modeling the background, the corners are detected for every frame and compound features are extracted. Data mining is performed by using Apriori algorithm as well as with the help of compound features extraction the action can be segregated in this video sequence. In the second phase, the same process is performed as well as analysis for new input frame and new pattern are updated for perfect recognition process. Due to Apriori algorithm, the processing delay is reduced and accuracy is improved.

Keywords: Feature extraction, action recognition, Apriori, SIFT.

I. INTRODUCTION

In most of computer vision application identifying moving objects in video sequence is a critical and fundamental task. Background subtraction is the common approach which identifies the moving object in a particular video frame that differs from background model. Good background subtraction algorithm should satisfy these challenges, it must be robust against changes in illumination, avoid detecting non-stationary background objects and internal background model should react immediately for the changes in background. Corner features, sometimes referred to as interest points, are image features characterized by their high intensity changes in the horizontal and vertical directions.

The four corner points are said to be good if the square object is present in the image. Shape and motion analysis can be performed by using these corners. Motion is ambiguous in nature at edges, since here corners are 2D image features ambiguity is not caused in motion analysis.

Within the object recognition, for feature selection learning has proven successful at building classifiers from large sets of possible features e.g. Boosting. Although various approaches similar to our approach have been applied to spatio-temporal activity domain, but those approaches is not good effective with the number of the features and also have issues with time alignment and scaling.

Our approach is based on extracting very low level features from the video sequence and these low level features are combined to form high-level, compound and spatio-temporal features. Data mining is used to accumulate the compound features using the data mining technique, Association rule, which efficiently

discovers frequently reoccurring combinations. To group SIFT descriptors for object recognition, association data mining was recently employed. We use it to build high level compound features from a noisy and over-complete set of low-level spatio and spatio-temporal features by which the action recognition can be done. We compare encoding only relative spatial offsets, which provides scale invariance, to the spatial grid proposed by Quack *et al.* and demonstrate that, due to increased scale invariance, higher performance is achieved. Learning is performed with only sequence class labels rather than full spatiotemporal segmentation. The resulting classifier is capable of both recognizing and localising activities in video. Furthermore, we demonstrate that efficient matching can be used to obtain real-time action recognition on video sequences.

II. RELATED WORK

The use of the spatial representation of local features within the field of object recognition has shown considerable success [3], [4], [5] and the temporal recognition of actions have been extended. A sparse selection of local interest points is used in many action recognition methods.. Sch'uldt *et al.* [6] and Dollar *et al.* [7] make use of sparse spatio-temporal features for the human action recognition. Sch'uldt takes a codebook and bag-of-words approach practical to single images to turnout a histogram of enlightening words or features for each action. Niebles and Fei-Fei [8] use a hierarchical model which can be characterized as a gathering of bags-of-words. Similarly Dollar take the bag-of-words approach but argue for an even sparser sampling of the interest points which improves the concert on the

same video sets. Conversely, the choice of feature used with such a sparse set of points is important. Scovanner *et al.* [9] extended the 2D SIFT descriptor [10] into three dimensions, by adding a further dimension to the orientation histogram that encodes temporal information and significantly outperforms the 2D version. To sulpt motion between frames, optical flow [11] [12] can be applied as was used by Laptev [6] in addition to a shape model to detect drinking and smoking actions. Yang Song *et al.* [13] use a triangular lattice of grouped point features to encode layout.

In the early work in action recognition was tested on moderately simple, single person, uniform background sequences [14], [22]. Laptev and Perez [15] expanded the ideas proposed by Ke *et al.* [16] to apply volumetric features to optical flow [17], [18]. Uemura *et al.* [15] used a motion model based on optical flow combined with SIFT feature correlation in order to accurately classify multiple actions on a sequence containing large motion and scale changes. A further idea that is being exploited to achieve success on complicated data sets is that of identifying context. Han *et al.* [19] and Marszalek *et al.* [20] learn the context of the environment in addition to the actual action. Han applies object recognition to learn relationships such as the number of objects and distance between them in order to boost a standard SIFT-based HoF/HoG [21] bag-of-words approach. Marszalek *et al.* [20] build on the previous work by Laptev *et al.* [21] by learning the context in which actions occur. Therefore, by detecting the scene in which the action is occurring, the action classification can be improved. The scene model is learned using 2D Harris corners with SIFT descriptors, while using the HoF and HoG descriptors of Laptev [21] to recognize the action.

The scale of the data sets in temporal-based action recognition directly lends itself to data mining algorithms, especially where only weak supervision is available. However, most previous applications of mining have been within the imaging field. Tesic *et al.* [28] used a data mining approach to find the spatial associations between classes of texture from aerial photos. While Quack *et al.* [10] applied Association rule data mining to object recognition by mining spatially grouped SIFT descriptors.

III. FEATURES

1. BACKGROUND MODELLING

An important part of tracking process in video sequences is background subtraction and removal for tracking motion objects. By removing static background we can accurately track the motion object and analyze their movements. Using spatial features of the image and remove background is the older technique for background removal where the new

technique is the use of temporal features which improves the background subtraction. Tracking motion objects in video sequences is a multiple process. This processes show in Fig. 1. For motion detection process background subtraction is a part by which we remove background, and only shadows, motion objects and motion noises remains. In the field of background subtraction there are many methods in that the older methods are based on object features such as color, intensity, edges, texture, etc. and relationship between frames is not considered in these methods.

Fig1. (a) Original frame (b) Finding the moving object (c) The image after background subtraction

2. EXTRACTING TEMPORAL 2D HARRIS INTEREST POINTS

Similar to sparse feature detectors, we fabricate our detection system lead to corner features. The underlying principle for using corners are they are easy to work out, largely invariant to both lighting and geometric transformation, and afford an over-complete feature set from which more complex compound features are constructed. Harris corner detector is the well known method to identify and locate interest points in the image. In our work, 2D Harris corner [24] method is used. Laptev and Lindeberg [25] proposed 3D corners as simple features in (x,y,t). These 3D corners are sparse, so instead, 2D corners are used independently by which the gradient interest points are found independently which yields information on spatial and temporal image changes is much denser than 3D Harris corners [25].

Fig2. 2D Harris corner detection on frames (a) Running (b) Boxing and Handclapping (c) Hug-person

Every corner feature has a dominant gradient orientation used to encode the feature type into single set of discrete corner orientations. Figure 1 shows the example corner detections on three frames. To detect corners at different scales the interest point detector was applied to the video sequences across scale space to overcome the effects of scale. This was achieved by successively 2x2 blocks averaging the image

frames. Table 1 shows the scale, image size and effective interest point patch sizes.

Table 1. Table showing the image and relative interest point patch sizes

Scale	1	2	3	4
Image size	160x120	80x60	40x30	20x15
Interest point size	3x3	6x6	24x24	48x48

III. SCALE-INVARIANT GROUPING

The key for object recognition is the scale invariant features which significantly improve the action recognition when modeled from temporal information independently. Quack *et al.* encoded the spatial layout of features by quantising the space around a feature into a grid and assigning features to one of those locations. In order to provide robustness to the scale, the size of the grid is dependent on the scale of the detected SIFT feature. This approach is difficult for achieving less descriptive interest points such as corners, so our approach is to define neighbourhoods centred upon the feature that encode the relative displacement in terms of angle rather than distance hence achieving scale invariance. To perform this, each detected interest points should form the centre of the neighbourhood.

Fig3. a) local feature descriptor, (b) interest points with local features shown as corners, (c) spatial and temporal encoding for local features and (d) for this interest point, local features concatenated into transition vector.

IV. DATA MINING

Data mining allows large amounts of data to be processed to identify any reoccurring patterns within the data in a computationally efficient manner. Association rule [25] mining is the well know mining algorithm. This was initially developed to analyze shopping done by customers in supermarkets, to the find the regularity of those customers in shopping behaviour. Using this millions of transactions the association rule is derived. An association rule, is a relationship of the form $\{A,B\} \Rightarrow C$, where A, B, and C are sets of items. A support and a confidence value plays important role because using this value only the belief of the rule is measured. From numerous transactions possible association rules are produced, to formulate the rules easily and quickly an efficient

algorithm is developed. Apriori algorithm developed by Agrawal is the popular algorithm to formulate the rules. The Apriori algorithm is a generative algorithm that uses a breadth-first, bottom-up strategy to explore item sets of increasing size, starting from single item-item sets and increasing the item set.

The frequency of an item set is related to the support and confidence for an association rule. An association rule of the form $A \Rightarrow B$ is evaluated by looking at the relative frequency of its antecedent and consequent parts, i.e., the item sets A and B. By using the statistical significance, support of the item set is measured, i.e., the probability that the item sets in the Transaction. For A, this is calculated as the size of the set of all T such that T is an element of D and A is a subset of T, normalized by the size of D. This can be formalized using set builder notation as,

$$\text{sup}(A) = \frac{|\{T \mid T \in D, A \subseteq T\}|}{|D|} \in \mathbb{R} \rightarrow [0, 1]. \quad (1)$$

The support of the rule $A \Rightarrow B$ is therefore,

$$\text{sup}(A \Rightarrow B) = \frac{|\{T \mid T \in D, (A \cup B) \subseteq T\}|}{|D|}, \quad (2)$$

and the statistical significance of the rule is measured. Then the confidence of the rule is calculated as,

$$\text{conf}(A \Rightarrow B) = \frac{\text{sup}(A \cup B)}{\text{sup}(A)} = \frac{|\{T \mid T \in D, (A \cup B) \subseteq T\}|}{|\{T \mid T \in D, A \subseteq T\}|}. \quad (3)$$

The probability of the joint occurrence of A and B is support, i.e., $P(A,B)$, while confidence is the conditional probability $P(B/A)$.

All generated association rules would be maintained and for other action classes the confidence would be used as a discrimination measure. This would be computationally infeasible due to the complete number of rules; therefore, generated rules are filtered using both support and confidence. At the initial level a single support value which is the lowest value is used throughout all of the stages of mining that is computationally feasible.

V. ACTION CLASSIFICATION

The frequently reoccurring distinctive and descriptive compound features for each class, are produced after the training phase is completed. By using the frequent item sets obtained from the data mining, the action classification of the unseen video sequence is done. Each transaction confidence is used to weight the matches and indicates that the Transaction T is distinctive compared to other classes using the high confidence. The use of the confidence ensures that if the transaction is matched with several classes, the confidence will provide a measure of the discrimination between those classes. If there are no matches found in the unlikely event, the model score is zero and the video would be classed as not containing any action.

Fig4. Action classification examples (a) boxing, (b) hand clapping, (c) hand waving, (d) jogging, (e) running, and (f) walking.

VI. EXPERIMENTAL RESULTS

To estimate the approach proposed, two different data sets were used. To illustrate the generalization method the focus of each data set is different. The well-known and popular KTH data set by Schüldt et al., to provide a comparison with the other existing techniques containing 6 different actions; boxing, hand-waving, hand-clapping, jogging, running and walking are shown in fig4. The simultaneous multi-action Multi-KTH data set, demonstrates detection of multiple actions in noisy scenes with background confusion and a moving camera.

To provide an additional challenge, the Multi-KTH data set was proposed. It consists of a single 753 frame long sequence, where multiple people perform the KTH actions simultaneously. To increase difficulty, there are large changes in scale, camera motions, and a non uniform background. Some frames from the sequence are shown in Fig.5.

Fig5. Example for the Multi-KTH data set, performing multiple KTH actions

Box	100	0	0	0	0	0
Clap	0	94	6	0	0	0
Wave	0	1	99	0	0	0
Jog	0	0	0	91	7	2
Run	0	0	0	10	89	1
Walk	0	0	0	0	6	94
	box	clap	wave	jog	Run	Walk

Fig6. Convolution matrix results for the KTH data set using training/test partition

To produce a Frequent Mined Configuration vector M the training sequences were used for each of the six actions containing compound features used to classify each of the test sequences. Figure 7(a) shows the classification confusion matrix using the scale invariant grid approach proposed within this paper, where good class separability is exhibited. The results show relatively little confusion compared to other approaches with minor confusion, this is consistent with previous approaches and it's due to the inherent similarity of the motion. In Figure 4(b) the experiments were repeated using a fixed size 4x4 grid similar, to investigate the importance of the spatial and temporal compounding of individual features.

	Spatio Temporal Grid							4x4 Spatial Grid					
Box	93	2	0	0	3	1	Box	84	2	14	0	0	0
Clap	14	84	0	1	0	1	Clap	1	98	1	0	0	0
Wave	2	0	92	1	0	4	Wave	15	0	85	0	0	0
Jog	3	0	0	87	1	6	Jog	0	0	0	92	15	3
Run	2	0	0	7	87	3	Run	0	0	0	15	83	0
Walk	0	0	0	0	4	96	Walk	0	0	0	3	0	97
	box	clap	wave	jog	run	Walk		box	clap	wave	jog	run	Walk

Fig7. The confusion of the data mined corner on the KTH dataset (a) Scale invariant spatial grouping (b) non-scale invariant spatial grouping.

Compared to other published methods Mined dense corners has a high classification accuracy, to select optimal low level features for discriminative classification.

Table 2. Comparison of Average precision with other techniques on KTH action recognition Dataset

Method	Average Precision
Nowozin <i>et al.</i> [23] Subseq Boost SVM	87.04%
Wong and Cipolla [24] Subspace SVM	86.60%
Niebles <i>et al.</i> [25] pLSA model	81.50%
Dollar <i>et al.</i> [5] Spat-Temp	81.20%
Schüldt <i>et al.</i> [1] SVM Split	71.71%
Ke <i>et al.</i> [3] Vol Boost	62.97%
Fixed Grid Mined Dense Corners	88.50%
Scale Invariant Mined Dense Corners	89.92%

The Multi-KTH data set is a more challenging version of the KTH data set. It has the same six actions and training video sequences, but the test sequence consists of multiple simultaneous actions, with significant camera motion. In the multi-KTH data set, the localization is performed to discriminate the multiple actions performed in the video sequence. For static actions the localization is generally focused on person's upper body and face, and is focused on

legs for the dynamic action, because the legs consists of the descriptive features.

Fig8. Example for multi-KTH localization Red- handclapping, Blue-boxing, Yellow-running, pink-walking and Green-hand waving

Figure 8 shows the various actions performed in the video sequence. To discriminate the multiple actions, different colors are used. Red-handclapping, Blue-boxing, Yellow-running, pink-walking and Green-hand waving. By using these colors the multiple actions performed in the video sequence are discriminated. The sample localization performed in our paper is shown in figure 9.

Fig9. Results from multi-KTH data set using localization (a) boxing, (b) handclapping, (c) running and (d) output frame

The main advantage of using the data mining technique is the speed of the learning patterns compared to the other machine language approaches. When compared to other techniques the simple 2D Harris corner detection has relatively a low computational cost. The spatial neighborhood grouping is fast to encode features which has limited neighborhood.

VII.CONCLUSION

This paper has presented an efficient solution to the problem of recognizing actions within video sequences with efficient learning of informative and descriptive local features for actions performed by humans at multiple scales. To form complex discriminative compounds of simple 2D Harris corners mined grouping corners are used which is fast. In a weakly supervised approach the frequently reoccurring patterns are learned by using data mining

approach where only class labels are required. Two different data sets have been tested, the Multi-KTH data set required multiple action localization and the KTH data set provides a comparison to other approaches. When tested on the popular KTH and multi-KTH dataset, notable results are obtained which do better than other state-of-the-art approaches. During training object segmentation is not required. To perform activity localization as well as classification the final classifiers can be used.

REFERENCES

- [1] Andrew Gilbert, John Illingworth and Richard Bowden, "Action Recognition Using Mined Hierarchical Compound Features", Transactions on Pattern Analysis and Machine Intelligence vol 33, no.5, may 2011.
- [2] Andrew Gilbert, John Illingworth, and Richard Bowden, "Scale Invariant Action Recognition Using Compound Features Mined from Dense Spatio-temporal Corners", CVSSP, University of Surrey, Guildford, GU2 7XH, England
- [3] T. Quack, V. Ferrari, B. Leibe, and L. VanGool, "Efficient Mining of Frequent and Distinctive Feature Configurations," Proc.
- [4] S. Lazebnik, C. Schmid, and J. Ponce, "Semi-Local Affine Parts for Object Recognition," Proc. BMVA British Machine Vision Conf., vol. II pp. 959-968, 2004.
- [5] J. Sivic and A. Zisserman, "Video Data Mining Using Configurations of Viewpoint Invariant Regions," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, vol. I, pp. 488-495, 2004.
- [6] Schudt, C., Laptev, I., Caputo, B.: Recognizing Human Actions: a Local SVM Approach. In: Proc. of International Conference on Pattern Recognition (ICPR 2004), vol. III, pp. 32-36 (2004).
- [7] Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior Recognition via Sparse Spatio-temporal Features. In: ICCCN 2005: Proceedings of the 14th International Conference on Computer Communications and Networks, pp. 65-72 (2005)
- [8] Niebles, J.C., Fei-Fei, L.: A Hierarchical Model of Shape and Appearance for Human Action Classification. In: Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2007) (2007)
- [9] Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proc. of MULTIMEDIA 2007, pp. 357-360 (2007)
- [10] Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 20, 91-110 (2003)
- [11] Dalal, N., Triggs, B., Schmid, C.: Human Detection using Oriented Histograms of Flow and Appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 428-441. Springer, Heidelberg (2006)
- [12] Lucas, B., Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision. In: Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI), pp. 674-679 (1998) Scale Invariant Action Recognition 233
- [13] Song, Y., Goncalves, L., Perona, P.: Unsupervised Learning of Human Motion. Transactions on Pattern Analysis and Machine Intelligence 25, 814-827 (2003)

-
- [14] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," Proc. Int'l Conf. Pattern Recognition, vol. 3, pp. 32-36, 2004.
 - [15] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp.2247-2253, Dec. 2007.
 - [16] I. Laptev and P. Pérez, "Retrieving Actions in Movies," Proc. IEEE Int'l Conf. Computer Vision, 2007.
 - [17] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient Visual Event Detection Using Volumetric Features," Proc. IEEE Int'l Conf. Computer Vision, 2005.
 - [18] N. Dalal, B. Triggs, and C. Schmid, "Human Detection Using Oriented Histograms of Flow and Appearance," Proc. European Conf. Computer Vision, vol. II, pp. 428-441, 2006.
 - [19] B. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," Proc. Seventh Int'l Joint Conf. Artificial Intelligence, pp. 674-679, 1998.
 - [20] M. Marszalek, I. Laptev, and C. Schmid, "Actions in Context," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2009.
 - [21] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 1-8, 2008.
 - [22] J. Tesic, S. Newsam, and B.S. Manjunath, "Mining Image Datasets Using Perceptual Association Rules," Proc. SIAM Int'l Conf. Data Mining, Workshop Mining Scientific and Eng. Datasets, p. 7177, 2003.
 - [23] T. Quack, V. Ferrari, B. Leibe, and L. VanGool, "Efficient Mining of Frequent and Distinctive Feature Configurations," Proc. 11th IEEE Int'l Conf. Computer Vision, 2007.
 - [24] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," Proc. Alvey Vision Conf., pp. 189-192, 1988.
 - [25] I. Laptev and T. Lindeberg, "Space-Time Interest Points," Proc. IEEE Int'l Conf. Computer Vision, pp. 432-439, 2003