Pattern Recognition Letters 33 (2012) 446–452

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Relevance feedback for real-world human action retrieval

Simon Jones [a], Ling Shao [a,*], Jianguo Zhang [b], Yan Liu [c]

[a] Department of Electronic & Electrical Engineering, The University of Sheffield, UK
[b] School of Computing, University of Dundee, UK
[c] Department of Computing, Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Article history:
Available online 11 May 2011

Keywords:
Content-based video retrieval
Relevance feedback
Human action recognition

ABSTRACT

Content-based video retrieval is an increasingly popular research field, in large part due to the quickly growing catalogue of multimedia data to be found online. Even though a large portion of this data concerns humans, however, retrieval of human actions has received relatively little attention. Presented in this paper is a video retrieval system that can be used to perform a content-based query on a large database of videos very efficiently. Furthermore, it is shown that by using ABRS-SVM, a technique for incorporating Relevance feedback (RF) on the search results, it is possible to quickly achieve useful results even when dealing with very complex human action queries, such as in Hollywood movies.

1. Introduction

The number of digital videos archived on the Internet grows daily at an enormous rate, on sites such as Youtube, Google Video, and countless others. It has become very easy and inexpensive for anyone to publish their own work on the Internet, through cheap digital video cameras and webcams. Despite this explosion in growth, however, the technology for accessing these videos has not been able to keep pace. Unlike text search engines, which directly search the content of a database of articles, current video search engines usually rely exclusively on textual metadata attached to the videos. These metadata are usually provided by the video's uploader and are, by nature, highly incomplete and are often inaccurate. Because of this, searches on such databases will give incomplete and inaccurate results.

To overcome these issues, much research has been done towards content-based video retrieval, an extension of Content-based Multimedia Information Retrieval (Lew et al., 2006) to the video domain. Here, the content of a video is searched directly, rather than arbitrary metadata. The content of most videos, however, is very noisy and contains a great deal of information, so knowing how that information can be extracted, and can be compactly represented, are both still open research questions.

Within this field, it is particularly important to address the topic of human actions, as humans are the subject of the majority of existing video media; however, retrieving realistic human actions poses a challenge to current information retrieval techniques. In addition to common computer vision problems such as lighting and varying viewpoints, the same human action can be performed in a great number of different ways – for instance, using different hands, performing the action from a different starting pose, or moving quickly or slowly. Furthermore, in a lot of video media the principal body parts involved in the action might be occluded or out-of-shot.

In this paper, we will apply a form of relevance feedback to the retrieval of human actions. This technique has previously been applied in the image domain, and we show that it can be extended to the video domain, even for very noisy datasets, such as those found on Youtube, or in Hollywood movies. In particular, we will be testing our algorithms on the *Hollywood* dataset (Laptev et al., 2008) of complex and realistic human actions. We will show that the use of Relevance Feedback (RF) can be used to greatly augment the accuracy of such a system after only a few iterations.

2. Related work

In this section we outline previous research in the fields of content-based multimedia retrieval, relevance feedback, and human action recognition, and show how our own work fits into this framework.

In order to recognise increasingly complex human actions, research has changed direction considerably over the past decade. Originally the focus was on the extraction of global features from videos – features describing the shape or appearance of the entire human body during the action. Such techniques typically rely on some form of background subtraction, and occasionally body part segmentation/localisation, as in Davis and Bobick (1997) and Shechtman and Irani (2005). Hidden Markov Models (HMMs) have been applied to classify human actions from global features with a great deal of accuracy, as introduced by in Yamato et al. (1992) and used in many subsequent works such as Feng and Perona (2002)

* Corresponding author.
E-mail address: ling.shao@sheffield.ac.uk (L. Shao).

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.05.001

and Weinland et al. (2007), due to their time-scale invariability. Nevertheless, global features do not perform well on noisy or crowded videos, and are sensitive to occlusions, multiple persons, moving backgrounds and differing camera viewpoints, making them unsuitable for recognition of actions in real-world scenarios.

To deal with this, much recent research has been made into local features. As the name suggests, local features are concerned with only small video patches within the overall action sequence; the points at which these video patches are extracted from the video are known as Space–Time Interest Points, or STIPs (Laptev, 2005). STIPs are incorporated into a model such as the Bag of Words model (Dollar et al., 2005), or a model containing structural information such as Spatio-temporal Shape Contexts (Shao and Du, 2009). These local features, while not as discriminative as global features in a very clean video, are far more robust against common problems such as partial occlusion, noise and differing viewpoints, making them suitable for a greater variety of applications. To detect STIPs, there are a variety of techniques, such as Dollár's method (Dollar et al., 2005), Laptev's method (Laptev, 2005) and Ning's method (Ning et al., 2007), and to describe the video patches, some more popular methods are Dollár's Gradient (Dollar et al., 2005), Laptev's HoG/HoF (Laptev et al., 2008) and 3D-SIFT (Scovanner et al., 2007).

More recent research on recognition of human actions includes that of Bregonzio et al. (1948) who suggest that global features can be extracted from a dense cloud of local features, providing a descriptor that successfully combines the discriminative power of holistic features with the robustness of local features. Additionally, work has been done to more closely approximate how the biological brain processes vision, such as in Jhuang et al. (2007) and Escobar et al. (2009). These methods have proven highly accurate against canonical datasets such as the KTH and Weizmann.

The majority of work in human action recognition to date has been done on simple datasets, such as the KTH Schuldt et al. (2004) and Blank et al. (2005); they are simple in that neither of these datasets are representative of real world human actions. Here, individual actors perform actions in a near identical fashion, from a fixed point of view, against a static background. The KTH adds more complexity by varying clothing and lighting, but it is still unrealistic. The Semantic Description of Human Activities 2010 (SDHA 2010) challenge introduced the UT datasets (Ryoo and Aggarwal, 2010; Chen et al., 2010; Ding et al., 2010), which incorporate human interaction and points of view that are common in real world surveillance, but they contain a static background and the actions are all performed orthogonally to the camera's viewpoint. Laptev et al. introduced a series of more complex datasets extracted from movies, such as the Hollywood dataset (Laptev et al., 2008). As the videos in these datasets were obtained from existing media, they are highly complex and present a real-world challenge.

While human action recognition has been an active research field for at least two decades now, human action *retrieval* – that is, content-based search of human actions – has to date not received much attention, though some recent efforts include Shao and Du (2009) and Jin and Shao (2010). Relevance Feedback, when applied to information retrieval, refers to the technique of iteratively incorporating user feedback on whether a set of results are relevant or irrelevant, to perform a new, more accurate query. It was first applied to textual information (Salton, 1971) but has more recently been shown to be effective when applied to image retrieval (Tong and Chang, 2001; Hong et al., 2000). As human action retrieval is relatively new, relevance feedback has not been much explored in this area, except in a recent paper by Jin and Shao (2010); however, only a very simple relevance feedback technique was used here, and the effect of applying multiple iterations of relevance feedback was not explored.

Most approaches for incorporating relevance feedback use SVMs, and attempt to learn the hyperplane separating relevant and irrelevant results. However, these techniques tend to perform poorly when there is only a limited number – or an asymmetric number – of positive and negative feedback samples provided by the user. There have been several attempts to overcome this. Tao et al. introduced an algorithm called Asymmetric Bagging and Random Subspace SVM (Tao et al., 2006), which uses several weak SVM classifiers to create a stable and accurate strong classifier, even in the presence of very few positive samples. Zhang et al. (2007) similarly used query expansion based on a set of soft, random sampling SVM classifiers. Other, more recent approaches to relevance feedback include Biased Discriminant Euclidean Embedding (Bian and Tao, 2010), Active Reranking for Web Image Search (Tian et al., 2010) and Negative Samples Analysis Method (Tao et al., 2007). So far, all of these techniques have only been applied to image datasets.

3. Methodology

This section presents our approach for information retrieval and relevance feedback applied to human actions in realistic scenarios. We wish to create a system that does the following:

Given an example human action video (henceforth known as the query), it will find all the most similar video sequences within a database of human action videos. The most similar video sequences will be ranked and presented to the user in order, and from these results the user will select some sequences which are relevant to the query (positive samples) and some irrelevant sequences (negative samples). Incorporating this feedback, the system will attempt to improve the results. The feedback stage can be repeated iteratively as many times as neccessary, until the results are satisfactory to the user. A diagram describing this system is shown in Fig. 1.

3.1. Representation of videos

In order to perform information retrieval, we create a Bag of Words model based on space–time interest points. For extraction of these STIPs, we use Dollár's method (Dollar et al., 2005), as the STIP detector given in this paper performed the best in the evaluation in Shao and Mattivi (2010). For description, we use the gradient + PCA method, as this popular method is quite accurate and simple to implement. While recent evaluation papers have shown other descriptors to be more discriminative, pure STIP accuracy is not the focus of our work, and does not affect the outcome of our work.

First, separable linear filters are applied to the video sequence, to get a response function for every (x,y,t) point. The response function is:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2, \tag{1}$$

where $g(x;y;\sigma)$ is a 2D Gaussian smoothing kernel applied on the spatial dimensions, and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally, and defined as follows:

$$h_{ev}(t;\tau,\omega) = -cos(2\pi t\omega)e^{-t^2/\tau^2}, \tag{2}$$

$$h_{od}(t;\tau,\omega) = -sin(2\pi t\omega)e^{-t^2/\tau^2}, \tag{3}$$

ω is treated as a constant in all cases, so the only variable parameters σ and τ correspond to the spatial and temporal scales respectively. For our experiments, σ was set to 2.4 and τ to 1.7.

To describe a located STIP, a spatio-temporal cuboid is extracted around it. The gradients along the x, y and t axes are calculated (after being smoothed at several scales), and are then concatenated into a single descriptor vector – this is known as the Gradient

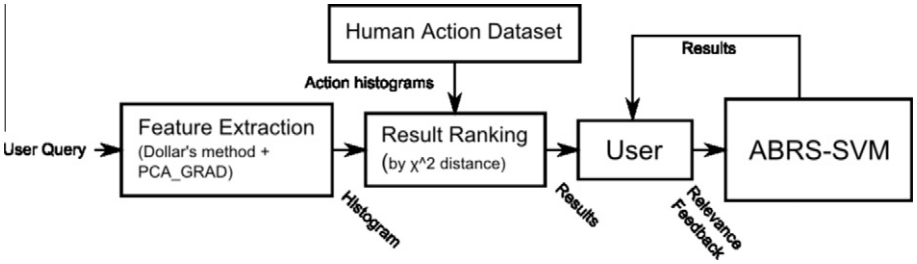

Fig. 1. A simplified diagram of a working information retrieval system.

method, from Dollar et al. (2005). The set of descriptors are later reduced in dimensionality by a round of PCA to capture 95% of the variation. In our experiments, the cuboid's dimensions were (17,17,13), for the x, y and t dimensions respectively, and smoothed at three different scales, resulting in a descriptor vector of length 11271, before being reduced by PCA.

Once all the features from a dataset have been extracted, we construct a video-word codebook from them, containing a vocabulary of k different types of feature. To achieve this, we perform k-means clustering on the feature descriptors for all of our dataset. Then, every feature in the dataset is assigned to the nearest video-word in Euclidean space, and for each individual video sequence, we construct an occurrence histogram of video-words, which shows how often each video-word appears in each sequence.

At this stage, every video sequence in the dataset, including the query video, can be represented as a histogram. We can use several metrics to determine the similarity between histograms, such as the χ^2 distance, the Euclidean distance, and the intersection; for our reported results, we used the χ^2 distance, as it gave the highest experimental accuracy.

Using this metric, we can determine the similarity of each video in the dataset to a given query video, and return the user an initial ranking of the most similar found videos. This method is efficient, as the histograms are much smaller than the video sequences they represent, and the comparison metric is simple to calculate. In our implementation, for a dataset with 449 videos, with a codebook of size 1000 (and therefore histograms of size 1000), it takes approximately 180 s to perform 100 queries.

3.2. Relevance feedback and ABRS-SVM

Once we have an initial ranking of videos against a query, the user iteratively provides feedback to get improved results. This feedback consists of a set of positive and negative examples from the top results, where positive examples are video sequences that are relevant to the search, and negative examples are considered irrelevant.

The Asymmetric Bagging and Random Subspace Support Vector Machine (ABRS-SVM) is a technique for incorporating relevance feedback used previously with some measure of success in image retrieval, as shown in Tao et al. (2006) and Li and Allinson (2009). It is designed to cope with three separate issues that often arise in relevance feedback systems:

- The number of feedback samples given is usually quite small, meaning an ordinary SVM will be unstable.
- There will often be more negative feedback than positive for very noisy/complex datasets, resulting in a biased hyperplane.
- The dimensionality of the feature vector is often much greater than the number of feedback samples, leading to overfitting.

In order to deal with the first two of these issues, we can use asymmetric bagging. This is random sampling with replacement on the set of negative examples S_{neg}, to produce n subsets $S_{b_{1..n}} \subset S_{neg}$, each the same size as the set of positive examples, S_{pos}. Then, T_s weak SVM classifiers are constructed, where the kth classifier uses $\{S_{pos}, S_{b_k}\}$ as its training set.

The random subspace method is employed to deal with the last issue – overfitting. Here, random sampling with replacement is applied to the feature space via bootstrapping, so that in every sample there is only a subset of the total features. Random sampling is performed T_f times and applied to all (positive and negative) feedback samples, resulting in T_f sets of feedback samples. Then T_f weak, linear SVM classifiers are constructed from each of these sets. This technique deals with the discrepancy between the high dimensionality of the feature vectors, and the small number of feedback samples.

These two algorithms are combined together to create the Asymmetric Bagging and Random Subspace SVM. First, asymmetric bagging is applied to generate T_s subsets of negative examples, and then the random subspace method is applied T_f times to each of these negative subsets, as well as the set of positive examples, so that there are a total of T_sT_f negative feedback sets and T_f positive feedback sets. These are then used to generate T_sT_f weak classifiers.

The weak classifiers resulting from ABRS-SVM are aggregated into a single strong classifier using the Bayes Sum Rule (BSR). BSR takes into account the relative informational value of each weak SVM, so that more accurate classifiers are given a stronger bias. It is defined as follows:

$$C^*(x) = argmax_k \left[(1-R)P(y_k) + \sum_{i=1}^{R} P(y_k|z_i) \right], \tag{4}$$

where $z_i (1 \leqslant i \leqslant R)$ is the ith classifier, $P(y_k)$ is the prior probability of the ith class, R is the number of classifiers, and $P(y_k|z_i)$ is defined as:

$$P(y_k|z_i) = 1/\{1 + exp(-|f_i(x)|)\}, \tag{5}$$

f_i is the output from the ith classifier.

4. Experiments

4.1. Datasets

To test the effect of relevance feedback, we initially used the KTH (Schuldt et al., 2004) and UCF Sports (Rodriguez et al., 2008) datasets, and then the Hollywood dataset of human actions (Laptev et al., 2008).

The KTH dataset is the canonical dataset currently used in human action recognition, consisting of 598 examples of 6 different simple, cyclical human actions, performed from near-identical, side-on viewpoints. Lighting, the actors, and clothing of the actors, however, are varied. The UCF Sports dataset consists of 150 examples of 13 categories of sports actions – the actors and settings once again varied, but the viewpoint within each category was consistent. We used these two datasets in order to show the particular challenge of the Hollywood dataset.

The Hollywood dataset consists of 449 video sequences taken from 32 popular Hollywood movies, and is split into 8 different classes of human action (for the specific actions see Fig. 2) Some of the video sequences are considerably longer than the actions within them; however, for the purpose of this paper we pre-processed the dataset to localise all of the actions within the video sequence, using the ground truths provided. The human actions in the Hollywood dataset are particularly challenging for state of the art algorithms to recognise, for several reasons, such as widely differing camera viewpoints, severe occlusion, different durations of activity, and different methods of performing the same activity. Fig. 3 illustrates these issues with still images taken from the dataset. Indeed, some of the examples here are far beyond the capabilities of current pattern recognition techniques to recognise, and would require the system to possess contextual knowledge about the world and humans in order to interpret them (for instance, see Fig. 3(c)). An additional difficulty with the Hollywood dataset is the unequal number of examples for each action class, and the sparsity of examples for one or two of the action classes. For example, the *Kiss* action is in a total of 100 video sequences, whereas *Handshake* is only in 39 video sequences. To compensate for this, we calculated the accuracy for a query as the percentage of correct items in the top $ceil(\frac{1}{5}I)$ results, where I is the number of items in the dataset with the same action class as the query. Before processing, every video in the Hollywood dataset was resized, maintaining the aspect ratio, to a height of 120 pixels – this is in part due to practical computational limitations, but also because the dataset favours close-up shots, so fine detail is unlikely to be important.

Prior to experimentation, we expect relevance feedback to be particularly effective in improving accuracy on the Hollywood; a single query cannot inform the system of intraclass variability, whereas further feedback examples allow us to model this to a limited extent.

4.2. Setup

In order to maximise utility of our datasets, we performed a set of round-robin tests. Each video sequence in turn was taken as the query, while the rest of the dataset was treated as the database from which to retrieve similar results. The number of features to extract from each video was calculated by dividing the total frames in the video by 5; we experimentally found this gave superior results to either feature strength thresholding, or a fixed maximum number of features, as both the length of examples and the strength of features vary greatly in the UCF Sports and Hollywood datasets.

For the initial query (before relevance feedback) we used the occurrence histogram and overlap distance, as described above, to rank the videos and get the top X results. After the initial ranking, the results were split into "positive" examples and "negative" examples, depending on whether they contained the same action class as the query video. To get the positive/negative feedback, we simulated user feedback using the ground truth data from the dataset. Thus, the positive set of examples was composed of any video sequence containing the same action class as the query video, and the negative set was the complement to the positive set. The first Y positive/negative results were then used in a round of relevance feedback with ABRS-SVM; by keeping Y low, we took into account that a real user will likely not be patient enough to provide more than a few feedback examples at a time.

Relevance feedback was performed iteratively for every query a total of nine times, recording the accuracy at every stage. We predetermined the optimal number of visual-words as roughly 1000 for the Hollywood dataset, and applied this to the KTH also. In our experiments, greater number of visual-words did not result in a significant accuracy increase, but did negatively affect running time. We varied X and Y to determine their effect on accuracy. In addition, we varied the parameters used in ABRS-SVM: T_s and T_f. The results for these are shown below.

All coding and experiments were conducted using Matlab, on a standard Core 2 Duo workstation with 4 GB of memory, running Windows 7. A single full experiment across the entire Hollywood dataset took approximately 3 h.

4.3. Results

As can be seen in the results shown in Fig. 4, it is clear that the relevance feedback aids retrieval performance considerably, reaching 93.2% accuracy for the KTH, 93.5% for UCF Sports, and 48.4% for the Hollywood after the ninth iteration. The large discrepancy in accuracy between Hollywood and the other two datasets is expected, as we outline in the dataset section above. Before relevance

Fig. 2. Action categories in the Hollywood dataset.

(a) An ambiguous Kiss action, requiring context. It could also be a HugPerson.

(b) StandUp obscured by a partially transparent curtain, accompanied by ego motion.

(c) Extreme occlusion and unusual viewpoint for a SitUp action.

Fig. 3. Examples of difficult to classify action sequences.

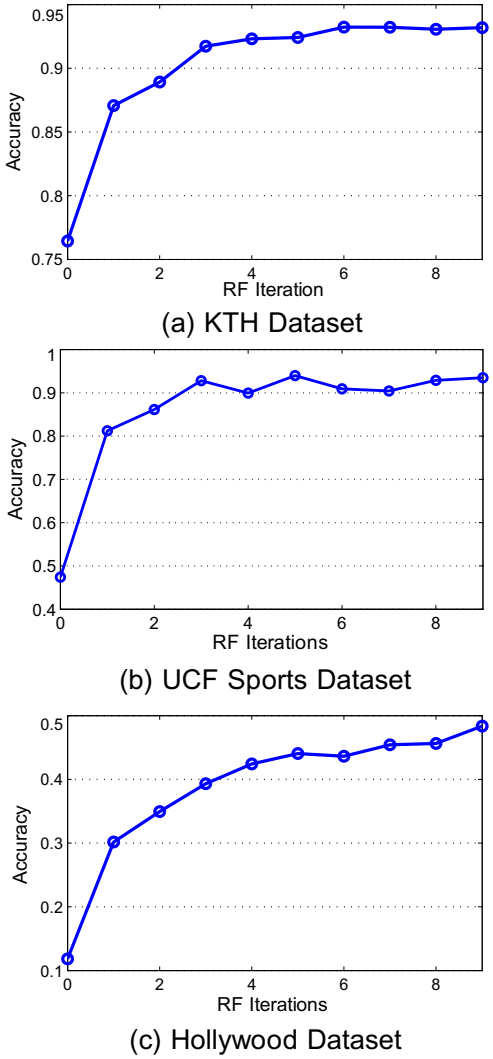

Fig. 4. Accuracy of the top $ceil(\frac{1}{5}I)$ results, over 9 rounds of RF on all datasets. $X = 20$, $Y = 5$, $T_s = 5$, $T_f = 5$.

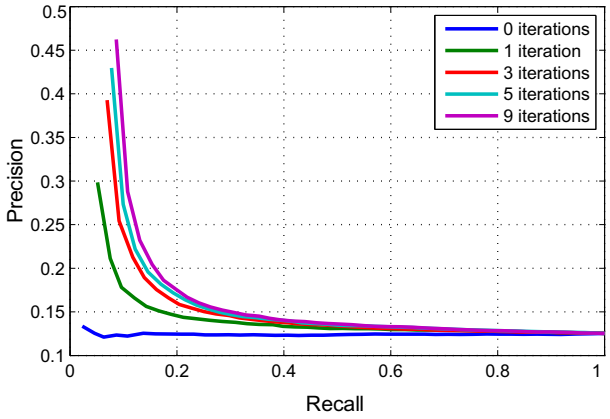

Fig. 5. Precision/recall curve for the Hollywood dataset, after different levels of relevance feedback.

feedback is applied, the Hollywood results are no better than would be expected by chance. We attribute this to the fact that a single video is not sufficient to model the huge intraclass variability of a Hollywood action class. After relevance feedback is applied, this changes considerably, and reaches 39.3% accuracy after only 3 iterations, and continues to rise thereafter.

Fig. 5 shows the precision/recall curve for the Hollywood dataset after each stage of relevance feedback. This shows improvement in search results for low recall – however, after about 20% recall, the precision starts to converge to chance for all levels of relevance feedback, demonstrating that there is a practical limit to how much feedback can improve results when the search terms are so noisy. Despite this, on a large enough dataset, applying our method would prove useful, as users are typically unlikely to look beyond the first few returned results.

Shown in Fig. 6 are the results for KTH and Hollywood broken down by action. We can see that certain types of action benefit considerably more than others from our method. In the KTH dataset, the handclapping action sees the most improvement, perhaps because the feedback helps learn the discriminative boundary between handclapping and handwaving. The kiss action in the Hollywood dataset improves the most after a single round of relevance feedback because of relatively low intraclass variability, but then hits an improvement ceiling early. Other, more variable actions, such as *AnswerPhone*, see a more gradual improvement over a large number of RF iterations.

We varied several parameters of the experiments on the Hollywood dataset. Fig. 7(a) and (b) show the effect of varying model parameters T_s and T_f. Clearly higher T_s and T_f are beneficial to accuracy, but we observed diminishing gains; additionally, higher T_s and T_f correspond to higher sampling rates on the sample space and the feature space, so this adversely affected the performance of the system. Therefore, there is a trade-off selection on the values of these parameters for real-world applications – for our experimental setup, we would recommend $T_s = 7$ and $T_f = 14$.

We also varied X – the number of results returned by the system – and Y – the number of positive and negative feedback samples given by the user – as shown in Fig. 7(c) and (d). Unsurprisingly, for larger X and Y, the improvement given by relevance feedback increased, reinforcing that the utility of the system is dependent on the quantity of feedback provided by the user.

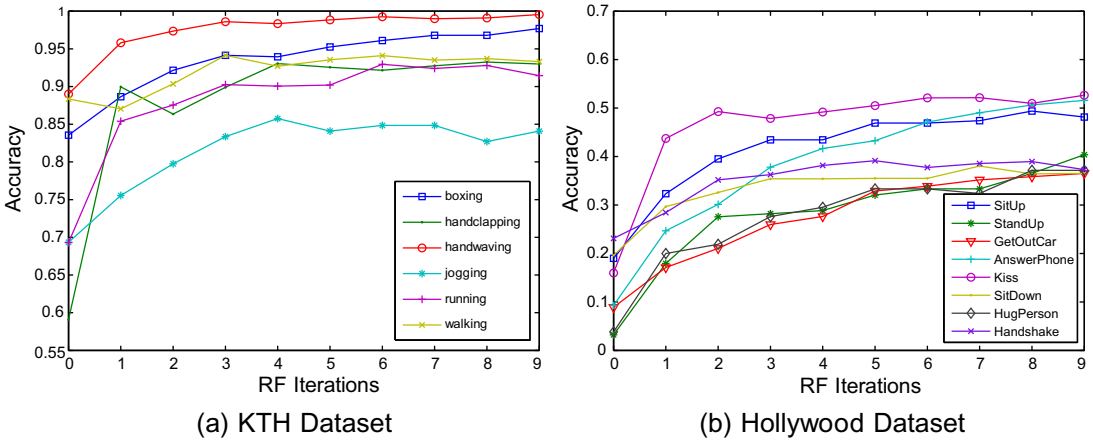

Fig. 6. Accuracy results for two of the datasets broken down by action. Parameters as in Fig. 4.

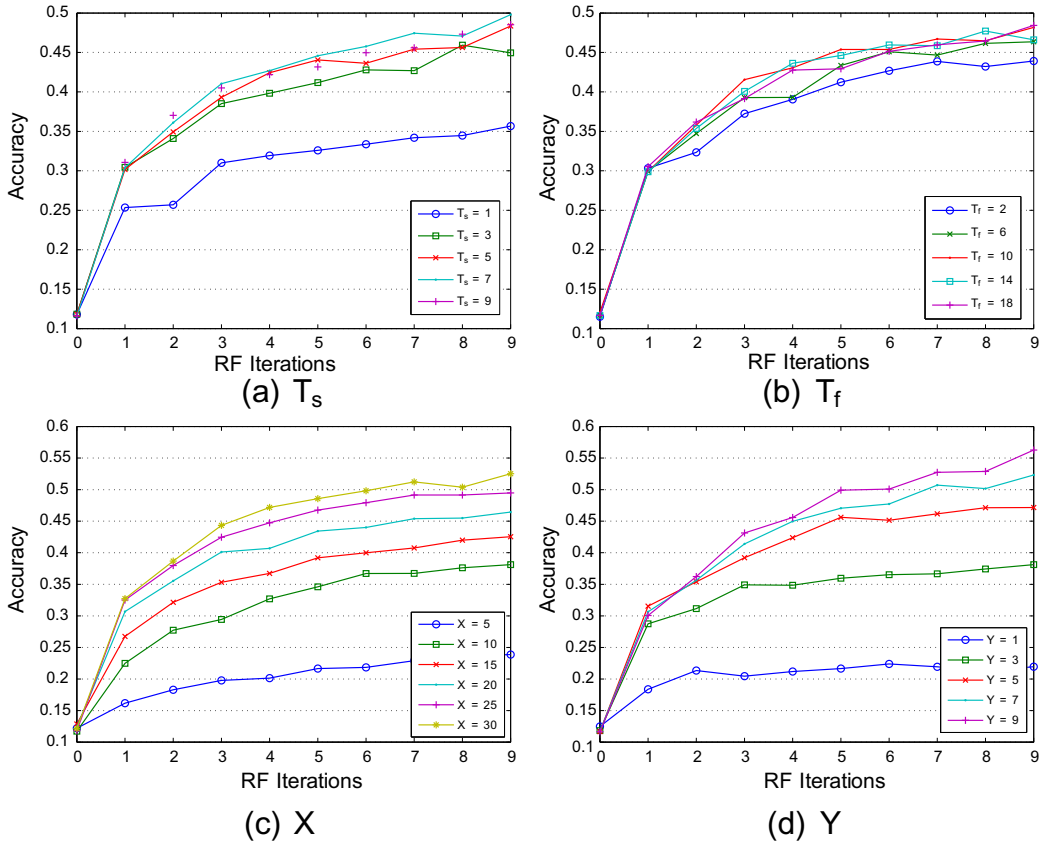

Fig. 7. Effect of varying various model parameters.

5. Conclusion

In this paper we have demonstrated the application of content-based information retrieval with relevance feedback in the video domain. In particular, we have focused on retrieving human actions from the Hollywood dataset, recognised as a particularly challenging dataset to work with, due to the very high intraclass variability. Differences in viewpoint, lighting, clothing and how the action is performed all confound the accuracy. Despite this difficulty, we have shown that it is possible to achieve, after only a few iterations of relevance feedback, significant improvements in accuracy of the search results, without semantic breakdown or cognitive understanding of the original query video.

While we have proved the efficacy of this method, however, such statistical techniques can only reach a certain level of accuracy, without further sophistication. Future work might include using more contextual information about scenes, or knowledge about the structure of the human body, in concert with relevance feedback, in order to further improve our ability to organise and search videos with complex human actions. In particular, this work could be combined with Marszalek et al.'s work on integrating object recognition with action recognition for enhanced results (Marszalek et al., 2009). Or, audio data from the scenes could be used to enhance recognition, as seen in Abdullah and Noah (2008) – for example, the distinctive sound of an opening car door could be used to enhance the accuracy of the GetOutCar action in

the Hollywood dataset. Finally, to further the practicality of this research, additional work could also be done on combined action retrieval and localisation, as real-world data are rarely conveniently annotated into short action sequences.

References

Abdullah, L.N., Noah, S.A.M., 2008. Integrating Audio Visual Data for Human Action Detection. In: Internat. Conf. on Computer Graphics Imaging and Visualization, pp. 242–246.

Bian, W., Tao, D., 2010. Biased discriminant euclidean embedding for content-based image retrieval. IEEE Trans. Image Process., 545–554.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R., 2005. Actions as Space-Time Shapes, in: Proc. IEEE Internat. Conf. on Computer Vision, p. 1395.

Bregonzio, M., Gong, S., Xiang, T., 2009. Recognising action as clouds of space-time interest points. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1948–1955.

Chen, C.-C., Ryoo, M.S., Aggarwal, J.K., 2010. UT-Tower Dataset: Aerial View Activity Classification Challenge, <http://cvrc.ece.utexas.edu/SDHA2010/Aerial_View_Activity.html>.

Davis, J.W., Bobick, A.F., 1997. The Representation and Recognition of Human Movement Using Temporal Templates. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 928.

Ding, C., Kamal, A., Denina, G., Nguyen, H., Ivers, A., Varda, B., Ravishankar, C., Bhanu, B., Roy-Chowdhury, A., 2010. Videoweb Activities Dataset, ICPR contest on Semantic Description of Human Activities (SDHA), <http://cvrc.ece.utexas.edu/SDHA2010/Wide_Area_Activity.html>.

Dollar, P., Rabaud, V., Cottrell, G., Belongie, S., 2005. Behavior Recognition via Sparse Spatio-Temporal Features. In: IEEE Internat. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72.

Escobar, M.-J., Masson, G., Vieville, T., Kornprobst, P., 2009. Action recognition using a bio-inspired feedforward spiking network. International Journal of Computer Vision, 284–301.

Feng, X., Perona, P., 2002. Human action recognition by sequence of movelet codewords. Int. Sympos. 3D Data Process. Vis. Transm., 717.

Hong, P., Tian, Q., Huang, T., 2000. Incorporate Support Vector Machines to Content-Based Image Retrieval with Relevance Feedback. In: Proc. IEEE Internat. Conf. on Image Processing, Vol. 3, pp. 750–753.

Jhuang, H., Serre, T., Wolf, L., Poggio, T., 2007. A biologically inspired system for action recognition. In: Proc. IEEE Internat. Conf. on Computer Vision, pp. 1–8.

Jin, R., Shao, L., 2010. Retrieving human actions using spatio-temporal features and relevance feedback. In: Shao, L., Shan, C., Luo, J., Etoh, M. (Eds.), Multimedia Interaction and Intelligent User Interfaces: Principles, Methods and Applications. Springer-Verlag.

Laptev, I., 2005. On space-time interest points. International Journal of Computer Vision 64 (2-3), 107–123.

Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B., 2008. Learning Realistic Human Actions From Movies. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1–8.

Lew, M., Sebe, N., Djeraba, C., Jain, R., 2006. Content-based multimedia information retrieval: state of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl. 2, 1–19.

Li, J., Allinson, N.M., 2009. Subspace learning-based dimensionality reduction in building recognition. Neurocomputing 73, 324–330.

Marszalek, M., Laptev, I., Schmid, C., 2009, Actions in context. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2929–2936.

Ning, H., Hu, Y., Huang, T., 2007. Searching Human Behaviors Using Spatial-Temporal Words. In: Proc. IEEE Internat. Conf. on Image Processing, pp. 337–340.

Rodriguez, M., Ahmed, J., Shah, M., 2008. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1–8.

Ryoo, M.S., Aggarwal, J.K., 2010, UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA), <http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html>.

Salton, G., 1971. The SMART Retrieval System—Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Schuldt, C., Laptev, I., Caputo, B., 2004. Recognizing Human Actions: A Local SVM Approach. In: Proc. IEEE Internat. Conf. on In Pattern Recognition, Vol. 3, pp. 32–36.

Scovanner, P., Ali, S., Shah, M., 2007. A 3-Dimensional SIFT Descriptor and its Application to Action Recognition. In: Proc. IEEE Internat. Conf. on Multimedia, pp. 357–360.

Shao, L., Du, Y., 2009. Spatio-temporal Shape Contexts for Human Action Retrieval. In: Proc. Internat. Workshop on Interactive Multimedia for Consumer Electronics, pp. 43–50.

Shao, L., Mattivi, R., 2010. Feature Detector and Descriptor Evaluation in Human Action Recognition. In: Proc. ACM Internat. Conf. on Image and Video Retrieval, pp. 477–484.

Shechtman, E., Irani, M., 2005. Space-Time Behavior Based Correlation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Vol. 1, pp. 405–412.

Tao, D., Tang, X., Li, X., Wu, X., 2006. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Trans. Pattern Anal. Machine Intell. 28, 1088–1099.

Tao, D., Li, X., Maybank, S., 2007. Negative samples analysis in relevance feedback. IEEE Trans. Knowl. Data Eng., 568–580.

Tian, X., Tao, D., Hua, X.-S., Wu, X., 2010. Active reranking for Web image search. IEEE Trans. Image Process., 805–820.

Tong, S., Chang, E., 2001. Support Vector Machine Active Learning for Image Retrieval. In: ACM Multimedia, pp. 107–118.

Weinland, D., Boyer, E., Ronfard, R., 2007. Action Recognition from Arbitrary Views using 3D Exemplars. In: Proc. IEEE Internat. Conf. Computer Vision, pp. 1–7.

Yamato, J., Ohya, J., Ishii, K., 1992. Recognizing Human Action in Time-Sequential Images using Hidden Markov Model. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 379–385.

Zhang, Z., Ji, R., Yao, H., Xu, P., Wang, J., 2007. Random Sampling SVM Based Soft Query Expansion for Image Retrieval. In: Proc. Internat. Conf. on Image and Graphics, pp. 805–809.