Spatio-temporal Video Retrieval by Animated Sketching

Steven Verstock[1], Olivier Janssens[2], Sofie Van Hoecke[2] and Rik Van de Walle[1]
[1]*Department of Electronics and Information Systems - Multimedia Lab, Ghent University – iMinds, Gaston Crommenlaan 8, bus 201, B-9050 Ledeberg-Ghent, Belgium*
[2]*ELIT Lab - University College West Flanders, Ghent University Association, Graaf Karel de Goedelaan 5, 8500 Kortrijk, Belgium*

Keywords: Query by Sketch, Video Retrieval, Motion History Images, Edge Histogram Descriptor, Animated Sketching.

Abstract: In order to improve content-based searching in digital video, this paper proposes a novel intuitive querying method based on animated sketching. By sketching two or more frames of the desired scene, users can intuitively find the video sequences they are looking for. To find the best match for the user input, the proposed algorithm generates the edge histogram descriptors of both the sketches' static background and its moving foreground objects. Based on these spatial descriptors, the set of videos is queried a first time to find video sequences in which similar background and foreground objects appear. This spatial filtering already results in sequences with similar scene characteristics as the sketch. However, further temporal analysis is needed to find the sequences in which the specific action, i.e. the sketched animation, occurs. This is done by matching the motion descriptors of the motion history images of the sketch and the video sequences. The sequences with the highest match are returned to the user. Experiments on a heterogeneous set of videos demonstrate that the system allows more intuitive video retrieval and yields appropriate query results, which match the sketches.

1 INTRODUCTION

Searching through a set of videos is still an open and challenging problem. Finding a video sequence which corresponds to what you have in mind is not always obvious. Today, to be able to retrieve the sequences in which a certain action is performed in a particular scene, one can either use text-based or example-based querying approaches (Petkovic and Jonker, 2004). However, both types of querying cope with some disadvantages, making it not always easy to find the desired sequence.

The main problem of text-based approaches, i.e. queries based on keywords and/or semantic annotations, is that it is not always possible to succinctly describe an image or a video sequence with words (Sclaroff et al., 1999). Furthermore, humans would probably describe the same content, i.e. the objects and the relations between them, with different words depending on their background. Another problem is that the same word can be used for totally different objects. As such, describing the query with words poses many problems. Additionally, the textual annotation of the set of video sequences, which is mostly done manual, is impractical, expensive and highly subjective (Brahmi and Ziou, 2004; Yang, 2004).

Exemplary-based video retrieval approaches, on the other hand, perform a query by using a description of (low-level) visual features (Aslandogan and Yu, 1999) or by using a similar video sequence (Lee et al., 2004), whose absence is usually the reason for a search. As is discussed in (Petkovic and Jonker, 2004), both examples of content-based video retrieval (CBVR) also yields poor results. The main problem is the multiple domain recall, i.e. similar features can occur in totally different content. Nevertheless, the advantage of CBVR is that the extraction of the features in the video database can be done (semi-)automatically, reducing the work a lot.

Recently, as a new way of exemplary-based video retrieval, query by sketch (QbS) is gaining importance. Although the number of QbS video retrieval approaches (Chang et al., 1998; Suma et al., 2008; Collomosse et al., 2009) is still limited and their usability is not always intuitive, sketch-based querying has already been used successfully for 'static' image retrieval (Eitz et al., 2009). By sketching the main feature lines of the scene, QbS image retrieval systems allow users to find images very intuitive. Since this way of querying is more connected to how humans memorize objects and their behavior, this paper further focus on this type of CBVR.

Verstockt S., Janssens O., Hoecke S. and Walle R..
Spatio-temporal Video Retrieval by Animated Sketching.
DOI: 10.5220/0004341607230728
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 723-728
ISBN: 978-989-8565-47-1

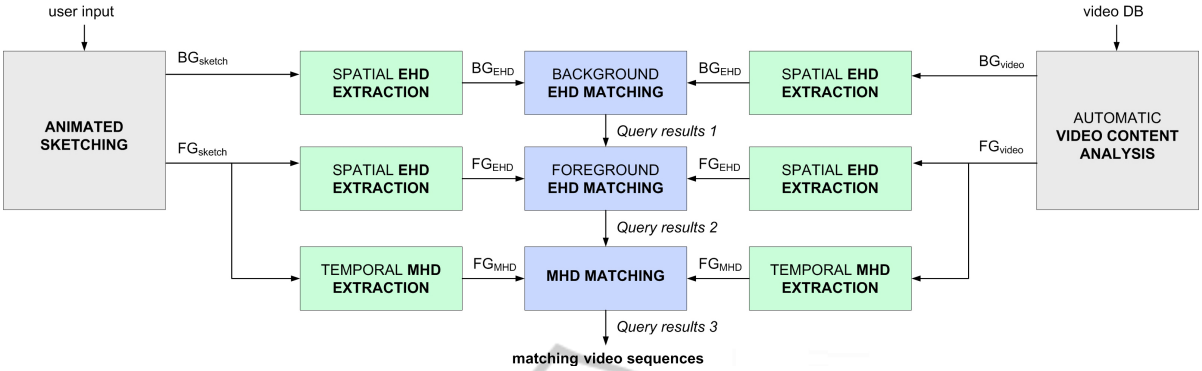

Figure 1: General scheme of QbS-based spatio-temporal video retrieval.

The main contribution of this paper is to extend QbS for 'dynamic' video retrieval. In order to do this, the spatial QbS is expanded with temporal query information by means of an animated sketch. This sketch lets the user intuitively describe the action he is interested in. An overview of the proposed system is given in Figure 1. First, we separate the BG_{sketch} and the FG_{sketch} of the animated sketch, i.e. the query created by the user. Then, the spatial edge histogram descriptors (EHD) of BG_{sketch} and FG_{sketch} are generated using a similar technique as in (Eitz et al., 2009). Next, the resulting BG_{EHD} and FG_{EHD} are matched against the EHDs of the video sequences in the video database (*video DB*). Important to mention is that the FG_{EHD} matching is only performed on the query results of the BG_{EHD} matching; only objects which appear in the 'desired' scene are further investigated. As soon as the *query results* 2 of the FG_{EHD} matching are available, the temporal analysis of foreground objects starts. In this last step, the novel temporal motion history descriptors (MHD) of FG_{sketch} and FG_{video} of *query results* 2 are matched. The matching video sequences, i.e. *query results* 3, are returned to the user.

The remainder of this paper is organized as follows. Section 2 gives a global description of the animated sketching. Next, Section 3 discusses the spatial EHD and the temporal MHD extraction. Subsequently, Section 4 proposes the spatio-temporal EHD-MHD matching. Then, Section 5 shows the evaluation results. Finally, Section 6 lists the conclusions.

2 ANIMATED SKETCHING

The creation of an animated sketch, shown in Figure 2, is performed in 3 or more steps, depending on the level of temporal detail and the number of moving objects the user wants to define. First, the user draws the background scene BG_{sketch} (step 1). When finished, this BG_{sketch} is disabled and made transparent, so that the user is not disturbed by it when drawing the foreground object(s). For each of the foreground objects FG_{sketch}, the user has to make at least two sketches: $FG_{sketch,n}$ (step 2) and $FG_{sketch,n+1}$ (step 3), so that the temporal description of the action/animation can be created. If more FG sketches are needed to describe the action, additional sketches can be made. When the user is ready with drawing the animated sketch, he can start the query operation.

The usability evaluation (Section 5) revealed that the proposed storyboard-way of animated sketching is found very intuitive and user work is minimal.

3 FEATURE EXTRACTION

In order to perform the sketch-based query, the system must be able to compare the rough sketched feature lines of the animated sketch with the full color image frames of the video sequences. However, due to their different image characteristics this is not quite straightforward. For that reason, i.e., to simplify the comparison, the proposed system extracts a spatial and a temporal descriptor. Both descriptors capture the essential properties of the common information in the sketch and the video sequences.

The proposed spatial EHD descriptor is based on the MPEG-7 edge histogram descriptor (Sikora, 2001) and the tensor descriptor proposed by (Eitz et al., 2009). Both focus on the orientation of the image gradients, which relates best to the direction of the sketched strokes. The temporal MHD descriptor, on the other hand, is a novel descriptor which extends the concept of motion history images (Bobick and Davis, 2001). The following subsections describe more in detail how both descriptors are extracted from the sketch and the video sequences. Subsequently, Section 4 discusses how the descriptors are matched.

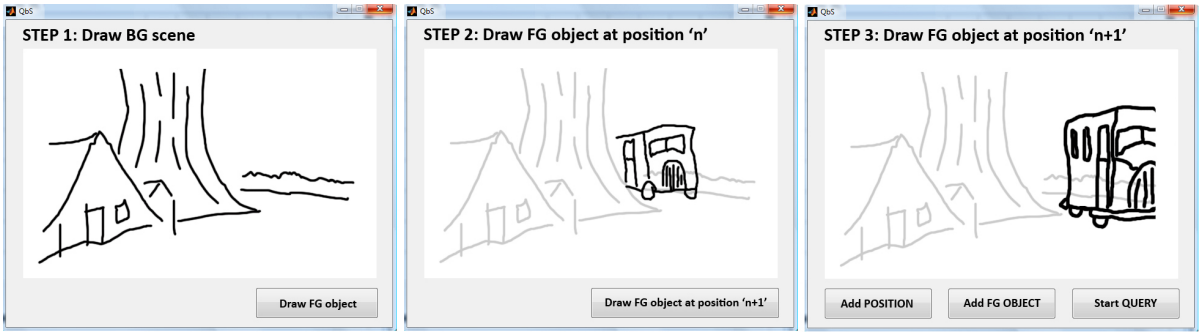

Figure 2: Graphical user interface for animated sketching.

3.1 Sketch EHD Extraction

To extract the sketch EHDs, we perform the algorithm shown in Figure 3. First, the input image I, i.e. BG_{sketch} or FG_{sketch}, is decomposed in 24x16 cells $C_{i,j}$. Next, for each cell, the gradient ∇J is calculated (Eq. 1), i.e. the first order derivative of $C_{i,j}$ in x and y direction. Then, based on this gradient, the orientation θ (Eq. 2) and magnitude $\|\nabla J\|$ (Eq. 3) of each pixel $[u,v]$ within $C_{i,j}$ are computed.

$$\nabla J = (J_x, J_y) = (\frac{\partial J}{\partial x}, \frac{\partial J}{\partial y}) \quad (1)$$

$$\theta = \tan^{-1}(\frac{J_y}{J_x}) \quad (2)$$

$$\|\nabla J\| = \sqrt{J_x^2 + J_y^2} \quad (3)$$

Subsequently, the gradient orientations are quantized into n_{bins} equally spaced bins $b_{i,j}(k)$ within $[0,\pi]$, with $k = 1 : n_{bins}$ ($\sim$ histogram analysis). This increases the flexibility of the system. By rounding the orientations it is not needed to have an exact match between the sketched lines and the detected lines in the video sequences; orientations of lines which are close to each other will belong to the same 'matching' bin. Note that negative gradients are also transformed into their positive equivalent, which discards the direction of intensity change, i.e. ambiguous sketch information. Furthermore, this also makes the system intensity invariant, improving the retrieval results.

During the orientation quantization we also do a thresholding on the gradient magnitudes. Based on the assumption that relatively stronger gradients are more likely to be sketched, only pixels with a high magnitude are further investigated. The thresholding itself is performed using Otsu's method (Otsu, 1979). Next, we count the number of 'strong' gradient pixels p, i.e. pixels [u,v] with $\|\nabla J_{u,v}\| \geq t$ (Eq. 4), within each orientation bin (Eq. 5). The gradient direction $\hat{k}_{i,j}$ of the bin $b_{i,j}(k)$ with the highest value is selected as a representative for the spatial 'structure' in $C_{i,j}$.

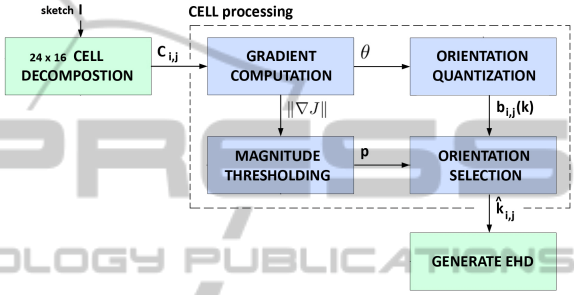

Figure 3: Sketch EHD extraction.

$$p[u,v] = \begin{cases} 1 & \text{if } \|\nabla J_{u,v}\| \geq t \\ 0 & \text{if } \|\nabla J_{u,v}\| < t \end{cases} \quad (4)$$

$$b_{i,j}(k) = \sum_{u,v \in C_{i,j}, \theta_{u,v} \in [k,k+1[} p[u,v] \quad (5)$$

Finally, the representative gradient directions $\hat{k}_{i,j}$ from all the cells $C_{i,j}$ are combined into the sketch edge histogram descriptor EHD. This sketch EHD is a 24x16 vector which contains the main gradients of the sketched image. By comparing this EHD with the $EHDs$ of the video sequences, the best spatial match(es) can be found, as explained in Section 4.

An example of the proposed sketch EHD extraction is shown in Figure 4. For each of the cells $C_{i,j}$ ($\sim$ Figure 2), the main gradient directions are shown. Note that empty cells, which do not contain sketched information, don't have a main gradient direction. In the EHD these cells are left empty and they will not be considered in the matching.

3.2 Video EHD Extraction

A schematic overview of the video EHD extraction is given in Figure 5. First, the set of video sequences is analyzed using a video content analyzer (VCA). The main goal of the VCA is to extract the different shots within the video sequences and generate their background BG_{video} and foreground objects FG_{video}.

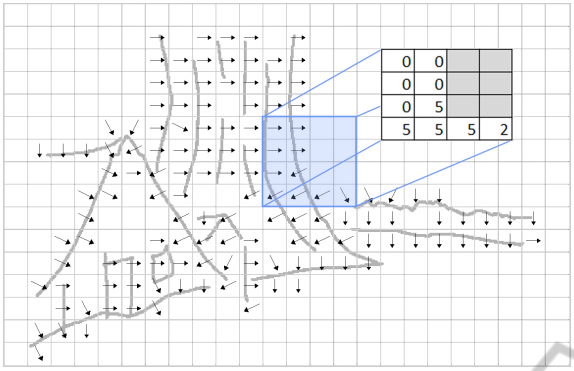

Figure 4: Example of BG_{sketch} EHD extraction.

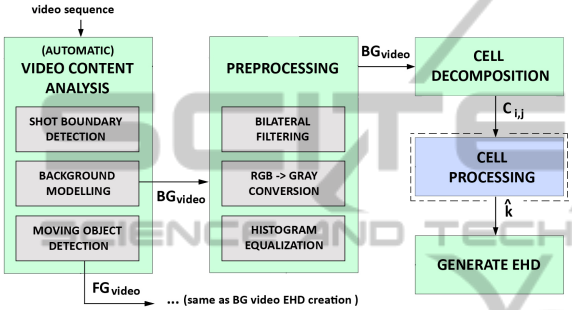

Figure 5: Video EHD extraction.

Subsequently, the BG_{video} and FG_{video} images are preprocessed for EHD extraction by bilateral filtering (Tomasi and Manduchi, 1998), RGB to gray conversion and histogram equalization, resulting in an 'appropriate' input for EHD extraction. Finally, the EHD of the preprocessed BG_{video} or FG_{video} is extracted. This EHD extraction consists of the same gradient-based cell processing and EHD generation as the sketch EHD extraction. An example of the BG_{video} EHD extraction is shown in Figure 6.

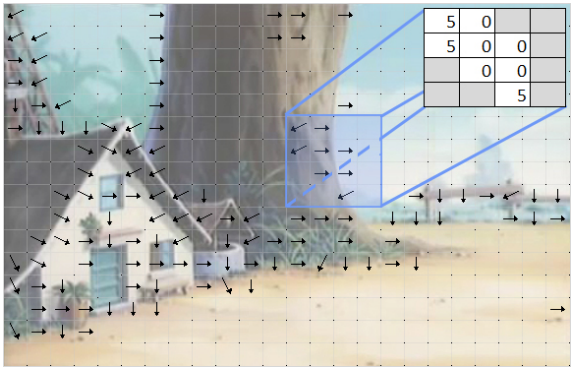

Figure 6: Example of BG_{video} EHD extraction.

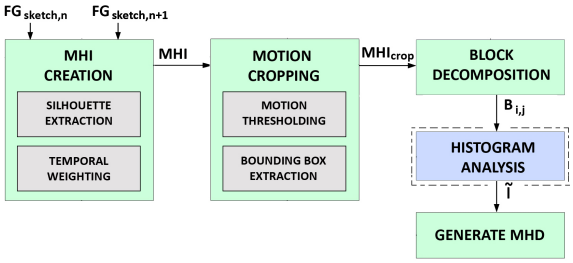

Figure 7: Sketch MHD extraction.

3.3 Sketch MHD Extraction

The temporal MHD descriptor is a novel descriptor which extends the concept of motion history images (MHI) (Bobick and Davis, 2001). To extract the MHD from a sketched foreground object FG_{sketch}, the algorithm shown in Figure 7 is followed. First, the MHI of the sketched FG object is created by combining the silhouettes of its 'temporal' sketches, i.e. $FG_{sketch,n}$ and $FG_{sketch,n+1}$ in our example. In the combined image, silhouettes at later positions appear more bright. This is done by weighting each of the silhouettes with its position. The creation of the silhouettes itself is done by boundary extraction and morphological filling. Subsequently, we crop out the bounding box around the motion part of the MHI, i.e. the non-black region. Then, the resulting crop is decomposed into 8×8 blocks $B_{i,j}$. Next, the intensity histogram for each block $B_{i,j}$ is analyzed to find the intensity value $\tilde{l}$ which occurs the most, i.e. the representative motion history value of that block. Finally, $\tilde{l}$ is stored at position [i,j] in the MHD vector. Figure 8 shows a clarifying example of the MHD creation.

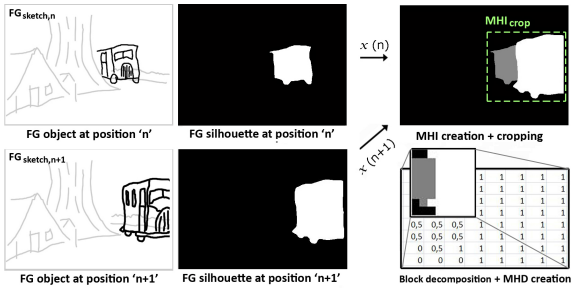

Figure 8: Example of sketch MHD extraction.

3.4 Video MHD Extraction

The video MHD extraction follows the same approach as the sketch MHD extraction, except its input is slightly different. As input it takes the background subtracted video frames. Then, for each of these frames it generates the silhouettes FG_{video} of its foreground object(s). Based on these silhouettes it then

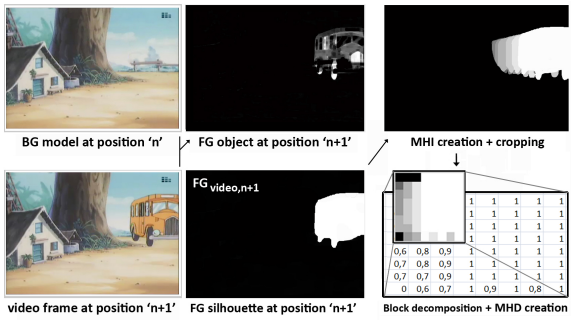

Figure 9: Example of video MHD extraction.

creates the MHI and the MHD in the same way as the algorithm shown in Figure 7. An example of the video MHI/MHD creation is shown in Figure 9.

4 FEATURE MATCHING

4.1 Spatial EHD Matching

The spatial EHD matching is similar to the matching process described in (Eitz et al., 2009). Both the BG and FG EHD matching are based on the computation of the distance between the sketch and the video 24x16 EHD vectors. The distance $d^{EHD}_{i,j}$ is defined as:

$$d^{EHD}_{i,j} = min(|\hat{k}_{i,j} - \hat{k}^*_{i,j}|, |(n_{bins} - \hat{k}_{i,j}) - \hat{k}^*_{i,j}|) \quad (6)$$

Important to mention is that when a non-empty sketch EHD cell does not have a matching video EHD cell, it gets a penalty score of 3, i.e. the maximal EHD distance. On the other hand, if a sketch EHD cell is empty, it is not taken in consideration.

By summation of the minimal circular Euclidean distance over all the non-empty EHD_{sketch} cells, the overall EHD distance d^{EHD} (Eq. 7) between the sketch and the video sequence is calculated.

$$d^{EHD} = \sum_i \sum_j d^{EHD}_{i,j} \quad (7)$$

The sequences for which the EHD distance of both the BG and FG objects is minimal, best match the spatial characteristics of the sketch and are selected as the spatial query results.

4.2 Temporal MHD Matching

After EHD matching, the temporal MHD matching of FG objects starts. First, the motion history values are row- and column-wise analyzed. For each MHD row and column we calculate a motion change value m, which is related to the number of positive $\#pos$ and negative $\#neg$ motion changes within the respective row or column:

$$m = \frac{\#pos - \#neg}{\#pos + \#neg} \quad (8)$$

Next, based on m and m^*, i.e. the motion change values of MHD_{sketch} and the MHD_{video} respectively, the distance d^{MHD} over their corresponding rows r and columns c is calculated:

$$d^{MHD} = \sum_r |m_r - m^*_r| + \sum_c |m_c - m^*_c| \quad (9)$$

Finally, the video sequences with minimal MHD distance are returned to the user as best match.

5 EVALUATION RESULTS

In order to objectively evaluate the proposed video retrieval system, the test setup shown in Figure 10 is followed. First, a random sequence is selected from the set of videos (composed of sequences from the Weizmann (Blank et al., 2005) and Hollywood datasets (Laptev et al., 2008)). After watching a sequence, the user is asked to draw an animated sketch of a representative action from this sequence. Next, the system queries the set of videos with the animated sketch. Finally, we check the position of the random sequence in the query results (Figure 11). The higher it occurs in the list, the better the system performs.

The objective measure used in our evaluation is the median query rank, which was also used in (Eitz et al., 2009). This measure gives a global idea of the overall performance of the system, as it 'averages' the results over all the random tests. Preliminary results show that the 'desired' sequence is retrieved in the top-10 video retrieval results with 87%. Although this can be regarded as a good result, some queries do not produce the expected results. The main problem is that the current system is not scale and position invariant. Preliminary tests on making the EHD scale and position invariant already show improvements, but further testing and fine-tuning is needed.

Besides the objective evaluation of the retrieval efficiency, the usability of the proposed system is also investigated using questionnaires. First questionnaire

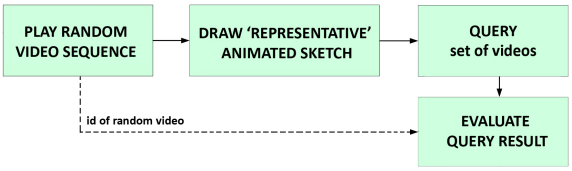

Figure 10: Evaluation test setup.

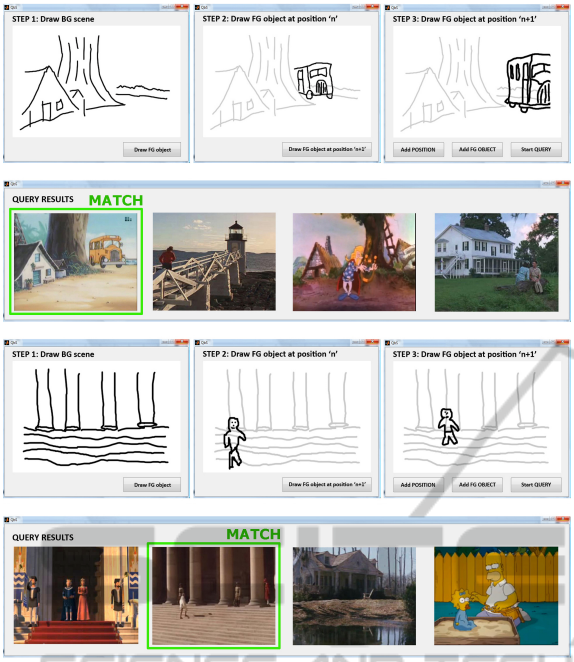

Figure 11: Exemplary animated sketching CBVR results.

results indicate that the proposed system has a positive usability score, although there are still a number of areas, such as the user interface, that could be improved. Multiple users also suggested to extend the system to a hybrid CBVR system that allows them to search by either text or sketch at any point of the searching process. This would provide them with a more flexible and powerful way of searching.

6 CONCLUSIONS

This paper proposes a novel intuitive querying method to improve content-based searching in digital video. By animated sketching users can easily define the spatial and temporal characteristics of the video sequence they are looking for. To find the best match for the user input, the proposed algorithm first compares the edge histogram descriptors of the BG and FG objects in the sketch and the set of video sequences. This spatial filtering already results in sequences with similar scene characteristics as the sketch. However, to find the sequences in which the specific sketched action occurs, this set of sequences is further queried by matching their motion history values to those of the sketch. The sequences with the highest match are returned to the user. Experiments show that the system yields appropriate query results.

REFERENCES

Aslandogan, A. Y. and Yu, C. T. (1999). Techniques and systems for image and video retrieval. *IEEE Transactions on knowledge and data engineering*, 11:56–63.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. *International Conference on Computer Vision*, pages 1395–1402.

Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates. *Transactions on Pattern Analysis and Machine Intelligence*, 23:257–267.

Brahmi, D. and Ziou, D. (2004). Improving cbir systems by integrating semantic features. *Canadian Conference on Computer and Robot Vision*, pages 233–240.

Chang, S. F., Chen, W., and Sundaram, H. (1998). Videoq: a fully automated video retrieval system using motion sketches. *IEEE Workshop on Applications of Computer Vision*.

Collomosse, J., McNeill, G., and Qian, Y. (2009). Storyboard sketches for content based video retrieval. *International Conference on Computer Vision*.

Eitz, M., Hildebrand, K., Boubekeur, T., and Alexa, M. (2009). A descriptor for large scale image retrieval based on sketched feature lines. *Eurographics Symposium on Sketch-Based Interfaces and Modeling*, pages 29–38.

Laptev, I., M., M. M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. *Computer Vision and Pattern Recognition*, pages 1–8.

Lee, A. J. T., Hong, R. W., and Chang, M. F. (2004). An approach to content-based video retrieval. *International Conference on Multimedia and Expo*, pages 273–276.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *Transactions on Systems, Man and Cybernetics*, 9:62–66.

Petkovic, M. and Jonker, W. (2004). *Content-Based Video Retrieval: A Database Perspective*, volume 1. Kluwer Academic Publishers, Norwell, MA, 1st edition.

Sclaroff, S., Cascia, M. L., Sethi, S., and Taycher, L. (1999). Unifying textual and visual cues for content-based image retrieval on the world wide web. *Vision and Image Understanding*, 75:86–89.

Sikora, T. (2001). The mpeg-7 visual standard for content description - an overview. *Transactions on Circuits and Systems for Video Technology*, 11:696–702.

Suma, E. A., Sinclair, C. W., Babbs, J., and Souvenir, R. (2008). A sketch-based approach for detecting common human actions. *International Symposium on Visual Computing*, pages 418–427.

Tomasi, C. and Manduchi, R. (1998). Bilateral filtering for gray and color images. *International Conference on Computer Vision*, pages 839–846.

Yang, C. C. (2004). Content-based image retrieval: a comparison between query by example and image browsing map approaches. *Journal of Information Science*, 30:254–267.