A Survey on Classification of Videos using Data Mining Techniques

[#1]Prashant Shambharkar, [#2]Nirmit Srivastava, [#3]Alok Yadav, [#4]Aseem Sharma, [#5]Ankit Katiyar
[#] *Department of Information Technology*
Krishna Institute of Engineering and Technology, Ghaziabad

ABSTRACT
Videos are in huge demand today. The internet is flooded with videos of all types like movie trailers, songs, security cameras etc. we can find so many genres but the only difficulty we face is the proper search of these videos. Sometimes we are irritated and get sick of the irrelevant search result. To sort out this difficulty we aim to classify videos on the basis of different attributes. Here in this paper we survey the video classification literature. Much work has been done in this field and much is awaited. We describe the general features chosen and summarize the research in this area. We conclude with ideas for further research.

Keywords
Video classification, Video databases, Genre classification

1. INTRODUCTION
Today there are more than 100 movies released every year in bollywood and if we add Hollywood the count goes above 300 and thanks to the Internet we can find all these movies online. The internet is flooded with videos of all kind, people can get their favorite movies at just one click but sometimes we get spurious results. The reason for this is the search technique used. We only search the database with the help of names but if we add some more attributes we can enhance the search technique. One more solution to this problem is Classification of Videos on the basis of genres. After classification the search space of the video is reduced and we can avoid getting redundant and spurious results and so classification of these videos is crucial for monitoring, indexing and retrieval purposes.

In this paper we focus on approaches to video classification, the method used and their results. Entertainment video, such as movies or sports, is the most popular domain for classification, but some classification efforts have focused on informational video (e.g., news or medical education). Many of the approaches incorporate cinematic principles or concepts from film theory. For example, horror movies tend to have low light levels while comedies are often well-lit. Motion might be a useful feature for identifying action movies, sports, or music videos; low amounts of motion are often present in drama. The way video segments transition from one to the next can affect mood. Cinematic principles apply to audio as well.

In [1], authors use visual disturbance and average shot length along with color, audio and cinematic principles to classify movie trailers. In [2] authors present a general comparison between several techniques using a general benchmarking system TrechVid, these techniques use all i.e. text, audio, and visual features, in this paper we can find a tabular comparison between various classification techniques. In [3] authors fused video signal along with the general features [9, 10, 11, 12, 13, 14] & [15] such as color, edge, texture and face for online videos. In [4] authors use motion and color features with Hidden Markov Models to classify summarized videos. In [5] authors use tags and focal points to classify videos on YouTube. In [6] authors focus on acoustic space, speaker instability and speech quality as audio based features to classify videos. SVM classifier is the base of this paper. In [7] authors use scene categorization, shot boundary analysis and bovw (Bag of visual words) to classify movie trailers. In [8] authors discuss the general introduction to video classification.

Previous work on video classification has two major limitations to be used on large-scale video databases. First, training and testing are generally performed on a controlled dataset. In a recent study, Zanetti et al. showed that most existing video categorization algorithms do not perform well on general web videos [6]. Furthermore, the sizes of the data-sets are relatively small when compared to the scale of online video services. Second, the algorithms treat each test video independently. We believe that online video services carry important cross-video signals that could be exploited to boost video classification performance. For instance, two videos that are uploaded by the same person, might share common information. Therefore, one should investigate whether the correlated information between multiple videos could be used for better video classification. In the literature, relatively little work address this problem. In [3], authors start with a small manually labeled training set and expand it using related YouTube videos.

In this paper we discuss the existing methods and procedures available and try to summarize them with some ideas for future research in this area.

2. GENERAL BACKGROUND
For the purpose of video classification, features are drawn from three modalities: text, audio, and visual. Regardless of which of these are used, there are some common approaches to classification.

While most of the research on video classification has the intent of classifying an entire video, some authors have focused on classifying segments of video such as identifying violent [1] [7] or scary [1] [7] scenes in a movie or distinguishing between different news segments within an entire news broad-cast [6].Most of the video classification experiments attempt to classify video into one of several broad categories, such as movie genre, but some authors have chosen to focus their efforts on more narrow tasks, such as identifying specific types of sports video among all video [3].

Many of the approaches incorporate cinematic principles or concepts from film theory. For example, horror movies tend to have low light levels while comedies are often well-lit. Motion might be a useful feature for identifying action movies, sports or music videos; low amounts of motion are often present in drama. The way video

segments transition from one to the next can affect mood [8]. Cinematic principles apply to audio as well. For example, certain types of music are chosen to produce specific feelings in the viewer [6].

In a review of the video classification literature, we found many of the standard classifiers, such as Bayesian, support vector machines (SVM), and neural networks. However, two methods for classification are particularly popular: Gaussian mixture models and hidden Markov models. Because of the iniquitousness of these two approaches, we provide some background on the methods here. Researchers who wish to use a probabilistic approach for modeling a distribution often choose to use the much studied Gaussian distribution. A Gaussian distribution, however, doesn't always model data well. One solution to this problem is to use a linear combination of Gaussian distributions, known as a Gaussian mixture model. An unknown probability distribution function p(x) can be represented by K Gaussian distributions such that

$$p(x) = \sum_{i=1}^{K} \pi_i \, N(x|\mu_i, \sum_i) \qquad \text{Eqn. (1)}$$

Where N (x|μi, ∑i) is the ith Gaussian distribution with mean μi and covariance ∑i. GMMs have been used for constructing complex probability distributions as well as clustering. The Hidden Markov model (HMM) is widely used for classifying sequential data. A video is a collection of features in which the order that the features appear is important; many authors chose to use HMMs in order to capture this temporal relationship. An HMM represents a set of states and the probabilities of making a transition from one state to another state. The typical usage in video classification is to train one HMM for each class. When presented with a test sequence of features, the sequence will be assigned to the class whose HMM can reproduce the sequence with the highest probability.

3. VARIOUS APPROACHES

Our approach is to breakdown movie trailers into frames and then uses these frames as "keyframes" to classify these trailers into various genres. To begin with the following approach is used for generation of keyframes.

A. SHOT DETECTION AND AVERAGE SHOT LENGTH

We explain the approach used in [1] and adapt it in our project. Authors in [1] classified the movie trailers into two basic categories action and non action movies and then further classifying the non action movies into comedy drama and horror with the help of lighting effects of movies based on the cinematic grammar principles. The algorithm used by the authors of [1] for the detection of shot boundaries using HSV color histogram intersection.

$$D(i) = \sum_{j \in allbins} \min(H_i(j) - H_{i-1}(j)) \qquad \text{Eqn. (2)}$$

Where *D*(*i*) represents the intersection of histograms *Hi* and *Hi¡*1 of frames *i* and *i ¡* 1 respectively. The shot change measure *S*(*i*) as

$$S(i) = D(i) - D(i-1) \qquad \text{Eqn. (3)}$$

Shot boundaries are detected by setting a threshold on S. For each shot, the middle frame within the shot boundary is picked as a *key frame*.

1) VISUAL DISTURBANCE IN THE SCENES: To find visual disturbance authors use an approach based on the structural tensor computation. The frames contained in a video clip can be thought of a volume obtained by combining all the frames in time.

This volume can be decomposed into a set of two 2D temporal slices, *I*(*x; t*) and *I*(*y; t*), also called horizontal and vertical slices respectively. Evaluation of the structure tensor of the slices as:

$$\Gamma = \begin{bmatrix} J_{xx} & J_{xt} \\ J_{xt} & J_{tt} \end{bmatrix} = \begin{bmatrix} \sum_w H_x^2 & \sum_w H_x H_t \\ \sum_w H_x H_t & \sum_w H_t^2 \end{bmatrix} \qquad \text{Eqn.(4)}$$

where *Hx* and *Ht* are the partial derivatives of *I*(*x; t*) along the spatial and temporal dimensions respectively, and *w* is the window of support. The direction of gray level change in *w*, *μ*, is expressed as:

$$R \begin{bmatrix} J_{xx} & J_{xt} \\ J_{xt} & J_{tt} \end{bmatrix} R^{T} = \begin{bmatrix} \Lambda_x & 0 \\ 0 & \Lambda_t \end{bmatrix} \qquad \text{Eqn. (5)}$$

where ¸*x* and ¸*y* are the eigen values and *R* is the rotation matrix. The angle of orientation *μ* is computed as:

$$\theta = \frac{1}{2} \tan^{-1} \frac{2J_{xt}}{J_{xx} - J_{tt}} \qquad \text{Eqn. (6)}$$

When there is no motion in a shot, *μ* is constant for all pixels.

In case of global motion the gray levels of all pixels in a row change in the same direction. This results in similar values of *μ*. However, in case of local motion, pixels that move independently will have different orientation. This can be used to identify each pixel in a column of a slice as a moving or a non-moving pixel. The density of *disturbance* is smaller for a *non-action* shot than that of an *action* shot.

We can observe that *action* movies have more local motion than a *drama* or a *horror* movie which results in a larger *visual disturbance*. Also shots in *action* movies change rapidly than in other genres, *drama* and *comedy* for example. The plot of visual *disturbance* against average shot length and uses a linear classifier to separate *action* movies from *non-action*.

After classification into action and non action authors in [1] use lighting to sub classify the trailers.

- *High-key lighting* The scene has an abundance of bright light with lesser contrast and the difference between the brightest light and the dimmest light is small. *High-key* scenes are usually happy or less dramatic. Many situation comedies also have high-key lighting.
- *Low-key lighting* The background and the part of the scene is generally predominantly dark with high contrast ratio. Low-key lighting being more dramatic are often used in Film Noir or horror films.

2) AUDIO ANALYSIS: In Hollywood movies, music and non literal sounds are often used to provide additional energy to the scene. The audio is always correlated with the scene. For example, fighting, explosions, etc. are mostly accompanied with a sudden change in the audio level. Therefore the energy in the audio track is computed as:

$$E = \sum_{i \in interval} (A_i)^2 \qquad \text{Eqn. (7)}$$

Where *Ai* is the audio sample indexed by time *i*. Interval was set to 50ms. We are interested in the instances where the energy in audio changes abruptly; therefore, we perform a peakiness test on the energy plot. A peak is *good* if it is sharp and deep.

B. CLASSIFICATION VIA SCENE CATEGORIZATION

The approach taken up by the authors in [7] is nearly similar to the approach in [1] because the domain of both is movie trailers. Cinematic principles are used for scene categorization and classification process.

Their approach to genre categorization is based on the hypothesis that scene categorization methods can be applied to a collection of temporally-ordered static key frames to yield an effective feature representation for classification. They explore this assumption by constructing such an intermediate representation of movie trailers. In their method, they decompose each trailer into a series of shots and perform scene categorization using state-of-the-art scene feature detectors and descriptors. These automatically-learned scene classes are then used as the "vocabulary" of movie trailers. Using the bag of visual words (bovw) model, each trailer can be represented as a temporally segmented 2D histogram of scene categories, which allows the calculation of trailer similarities.

1) SHOT BOUNDARY DETECTION: The first step of their approach decomposes a trailer into a series of shots using the shot boundary detection algorithm described in [1]. A trailer is first converted into its n frames. For each frame i, generate a histogram Hi of its HSV color space representation, with bin dimension 8, 4, and 4 for the hue, saturation, and value components, respectively.

2) SCENE CATEGORIZATION: The shot boundary detection step converts a set of trailers into a collection of shot keyframes kij, where i is the trailer index and j is the shot sequence index. The scene features from keyframes can now be analyzed using several state-of-the-art feature detectors and descriptors. In [7], they choose GIST, CENTRIST, and a variant call W-CENTRIST.

3) GIST: The GIST model produces a single, holistic feature descriptor for a given image, which attempts to encode semantic information describing characteristics of the image scene.

4) CENTRIST: CENTRIST, the CENsus TRansform hISTogram, is a visual descriptor developed for recognizing the semantic category of natural scenes and indoor environments, e.g. forests, coasts, streets, bedrooms, living rooms, etc. It has been shown that CENTRIST produces outstanding results for the place and scene recognition task when used within the bovw framework.

5) W-CENTRIST: Both GIST and CENTRIST models discard color information when generating descriptors. However, authors think color plays an important role in conveying the mood of a scene. Therefore, they have devised W-CENTRIST, a variant of the CENTRIST descriptor that captures color information in addition to intensity statistics.

Given the shot codebook, each trailer can be represented by the bovw model as a histogram of the shot classes that appear within it. These features are used to develop histogram which is further used for classification of movie trailers.

C. FUSING THE CROSS-VIDEO SIGNALS

Cross-video signals are extracted from a group of videos which are *related* to the test video. Two videos can be related via several ways. In this study, four cross-video relation sources are considered:

Two videos are assumed to be

• *Co-Browsed* if more than a certain number of users browsed both of them within a session.

• *Co-Uploaded* if the same user uploaded both videos within a certain time window.

• *Co-Commented* if more than a certain number of users commented on both of them.

• *Co-Queried* if more than a certain number of users clicked on both videos in response to the same queries.

Given a test video and the sets of related videos, the classifier *C* is applied on all of these videos independently to obtain the classification scores.

The classification score of the test video is denoted by *po*. Then, the median classification score is computed for each set of related videos. The median scores for co-browsed, co-uploaded, co-commented and co-queried videos are denoted by *pb*, *pu*, *pc*, *pq*, respectively. The final classification score, *pf* is obtained by taking the weighted sum of these scores. For ease of notation, the score vector is represented with

$$p = [po, pb, pu, pc, pq]T . \qquad \text{eqn.(8)}$$

Similarly the weight vector is represented with

$$w = [wo, wb, wu, wc, wq]T . \qquad \text{eqn. (9)}$$

$$pf = wTp \qquad \text{eqn. (10)}$$

Given the training instances (with labels *{yi}*) and corresponding classification score vectors *{*p*i}*, the task is to learn the weight vector which minimizes the classification error

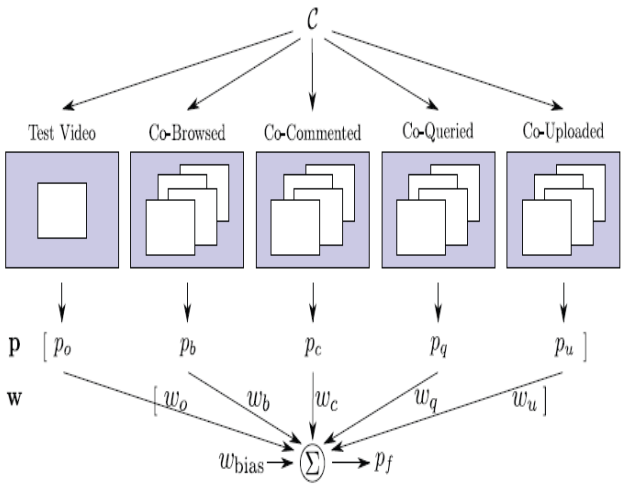

Fig. (1) Classification [3]

D. VIDEO CLASSIFICATION BASED ON HMM

Authors in [4] propose a hidden Markov model based method for video topic classification using visual and temporal features. They classify videos into four topic categories: news report, commercials, basketball game and football game.

1) Hidden Markov Model: In an HMM, there are a finite number of states and the HMM is always in one of those states. At each clock time, it enters a new state based on a transition probability distribution depending on the previous state. After a transition is made, an output symbol is generated based on a probability distribution, depending on the current state.

A HMM is always represented by l = (*A*, *B*,p) . Four HMMs are constructed corresponding to news, commercials, football game, and basketball game, respectively.

2) HMM Process: The HMM process consists of two phases, *viz.* training and classification

- *Training* - The HMM training step is essentially to create a set of hidden states *Q* and a state transition probability matrix *A* for each video topic category. The state transition probability matrix includes the probabilities of moving from one hidden state to another. There are at most M2 (M is the number of states) transitions among the hidden states. Since each of keyframe clusters obtained from the above step corresponds to a hidden state and each keyframe corresponds to a set of frames, calculate the probabilities based on the number of frames falling into these clusters and the number of frames temporally transiting between clusters.
- *Classification* - In the classification phase, given a target video, authors make an observation sequence and feed it into every HMMs. By evaluating the probability for each HMM, the target video is assigned into a topic category with the highest probability of the HMM. For the observation sequence of the target video, first summarize the target video and extract a set of keyframes in time order and take these keyframes as observation symbols. Then build a *temporal and keyframe-based summarized video sequence* (TSV) that is replicating each keyframe a number of times equaling the number of frames represented by the keyframe in the video sequence, and order these keyframes by time. In this way, a temporal feature can be maintained in the resulting sequence.

Authors use the *Forward* algorithm to calculate a probability for each HMM, and thus choose the video type with the most probable HMM.

E. CLASSIFICATION USING TAGS AND FOCAL POINTS

According to the author in [5] tags are the new valuable source information in the web. The multi-media content has been heavily tagged by the owners and the viewers. Thus, Tags represent a social classification of the content and at the same time it also adds to its semantics. The Authors uses tags and focal points to classify videos present on YouTube. The author focuses on two main points, i.e. Tags and focal points. The profile of a tag and focal point can be understood with the help of the following tables respectively.

Table 1 - Profile Of a "Tag"	
Definition:	
Loose	Tags are the keywords or labels that assigned to an object.
Strong	Tags are textual meta-data, i.e. text data about data, where the actual data can be a multi-media object or a text object.
Properties:	• Assignment of Tags to an object is highly user centric. • They are loosely structured, i.e. normally they are represented in the form of *words* separated by some delimiter like white spaces or comma. • The world of tags is inherently *flat*. There is no hierarchy in the relationship between tags that describe an object.
Benefit:	• Ideally tags represent the principal semantic component of an object. • The presence of linguistics in Tagging provides stability to otherwise complex environment [3]. • Tags are extensively used and play an important role in information retrieval on large tag based databases like YouTube.
Issues:	• They are loosely structured. • People tend to use same tags for different documents. • People also to tend to use semantically un-correlated tags for the same document.

Fig (2) Profile of a "Tag" [5]

Table 2 - Profile Of "Focal Points"	
Definition:	A focal point, also called Schelling Point [A concept introduced by Thomas Schelling in his book *The Strategy of Conflict*], is a solution that people will tend to use in absence of communication, because it seems natural, obvious and relevant to them.[6] Thomas Shelling, in the book "*The Strategy of Conflict*", describes "focal point[s] for each person's expectation of what the other expects him to expect to be expected to do"[5]
Example:	A classic example is the solution most people choose when asked to divide $100 into two piles, of any sizes; they should attempt only to match the expected choice of some other, unseen player. Usually, people create two piles of $50 each, and that is what Schelling dubbed a focal point.[4]

Fig (3). Profile of "Focal Points" [5]

[5] Identifies four general kinds of focal points that people tend to use:

- *Centrality*: People tend to give prominence to a choice that is at the center of set of choices.
- *Extremeness*: People tend to give prominence to the choices that are extreme to other choices. Choosing a red color square a 3x3 grid, where all other squares of light color.
- *Firstness*: People tend to give prominence to the choices that appear first in the set of choices. The example for this case is the tags for some content. The first tag usually represents the most important idea about the video.

Video Title: Apple WWDC 2006-Windows Vista Copies Mac OS X
Video Link: http://www.youtube.com/watch?v=N-2C2gb6ws8
Video tags: Apple Macworld Keynote Mac Macintosh Steve Jobs Bill Gates Windows OSX Computer iPod imac ibook power macbook pro vista. The video predominantly deals with Apple

computers; under this assumption the author of this video gave more priority to the tags like apple and MacWorld that have high co-relation with the video by placing them first!

- *Singularity*: People tend to give prominence to the choices that are unique as compared to the other set of choices. Choosing a red colored square on a 3x3 grid, where all other squares are of blue color.

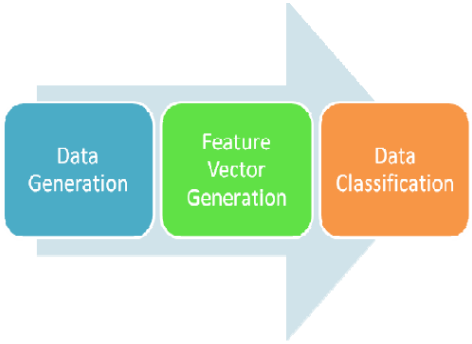

Fig (4) Data classification [5]

F. Robust Audio-based Classification of Video Genre

Authors in [6] use the audio features to classify videos. Their approach relies on the combination of low and high level audio features. The genre identification is performed on these features by using a SVM classifier.

They have selected 5 categories that are commonly targeted by video genre classification tasks: news, movies, cartoons, music and commercials.

1) System Overview: The system proposed is a 2-level architecture, where the first level consists in features extraction that are combined at the second level.

- *Acoustic space*: this is the most frequently used descriptor for categorization by audio only. The general idea is to distinguish genres by statistical modeling of their cepstral patterns.
- *Speaker interactivity*: video genre could differ from their interactivity profile; for example, speaker turns and speaking time are probably different between cartoons and news. The number of speakers and how they communicate together could differ according to the genre. For example, there is usually only one main speaker in news, when cartoons or movies contain generally many speakers with highly variable speaking times and speaker turns. The interactivity features aim at represent these speaker-related profiles.
- *Speech quality*: most of the speech-quality related features rely on speech recognition methods; we estimate the quality of speech contents by acoustic analysis and by an a posteriori evaluation of a speech recognition process that is applied to the speech segments. The basic idea is that the speech quality could provide some relevant information about genres. For example, speech is usually clean in news, whose the linguistic domain is well covered by speech recognition systems, since linguistic domain of commercials may be unexpected due to the product specificities and speaking styles.
- *Instability*: these features represent the regularity of the acoustic stream in the time domain.

Considering all the above features a SVM classifier is trained and classification is done in the desired classes.

4. CONCLUSION

We have reviewed the video classification literature and found that a large variety of approaches have been explored. Features are drawn from three modalities: text, audio, and visual. The majority of the literature describes approaches that utilize features from a single modality.

While much has been done, there are still many researches opportunities in automatic video classification and the related field of video indexing. Only a few of the papers that we reviewed attempted to perform classification at the shot or scene level. Being able to classify at the shot or scene level has many applications, such as content filtering (e.g. identifying violent scenes), identification of important scenes, and video summarization. This would also be useful in subdividing genre, such as creating a category of action movies that include car chases.

In [1] authors took 19 movies under consideration and fairly classified each of them into respective classes.

In [3] with the help of cross video signals authors show up to 4.5% absolute equal error rate (17% relative) improvement over the baseline on four video classification problems. Their future work will include extending the system for multi-class classification.

In [4] authors have described a video content classifier based on HMM using chromaticity signatures from video summarization and their temporal relationship. The video characterization and summarization method represents the video as a series of compressed chromaticity signatures. The HMM process uses these signatures and takes advantage of the temporal feature to train HMMs and evaluate the probability of the given video being in one of the four categories of TV programs. In [5] authors show that with the help of tags and focal points we can clearly boost the run performance. In [7] authors have presented a framework for automatic classification of film genres using features from scene analysis. They have demonstrated that a temporally-structured feature based on this intermediate level representation of scenes can help improve the classification performance over the use of low-level visual features alone. In the future, they will build upon their static scene analysis to include scene dynamics, such as action recognition and camera movement estimation, to help achieve higher-level dynamic scene understanding.

REFERENCES

[1] Zeeshan Rasheed ,Mubarak Shah,"Movie Genre Classification By Exploiting Audio-Visual Features Of Previews".

[2] Darin Brezeale and Diane J. Cook,"Automatic Video Classification: A Survey of theLiterature".

[3] Mehmet Emre Sargin, Hrishikesh Aradhye,"Boosting Video Classification Using Cross-Video Signals".

[4] Cheng Lu, Mark S. Drew, James AU**,"** Classification Of Summarized Videos Using Hidden Markov Models On Compressed Chromaticity Signatures".

[5] Ankur Satyendrakumar Sharma, Mohamed Elidrisi," Classification of Multi-Media Content (Video's on YouTube) Using Tags and Focal Points".

[6] Mickael Rouvier, Georges Linar`es, Driss Matrouf,"Robust Audio-based Classification of Video Genre".

[7] Howard Zhou, Tucker Hermans, Asmita V. Karandikar, James M. Rehg,"Movie Genre Classification via Scene Categorization ".

[8] Matt Roach, John Mason, Li-Qun Xu, Fred Stentiford,"Recent Trends In Video Analysis: A Taxonomy Of Video Classification Problems".

[9] Assfalg, J., Bertini, M., Bimbo, A. D., Nunziati, W., and Pala, P. "Soccer highlights detection and recognition using HMMs". IEEE International Conference on Multimedia and Expo 2002.

[10] L. J. Latecki and D. de Wildt, "Automatic Recognition of Unpredictable Events in Videos", ICPR. Vol. 2. 2002.

[11] Aznaveh, M.M., Mirzaei, H., Roshan, E. and Saraee, M.: "A new color based method for skin detection using RGB vector space", Human System Interactions Conference (2008) 932-935.

[12]Tsishkou,D., Hammami, M. and Chen, L.: Face Detection in Video Using Combined Data-mining and Histogram based Skin-color Model, IEEE Third International Symposium on Image and Signal Processing and Analysis, Rome, Italy (2003) 500-503.

[13] Sigal, L., Sclaroff, S. and Athitsos, V.: Estimation and Prediction of involving Color Distributions for Skin Segmentation under Varying Illumination, Proceedings of IEEE Conf. Computer Vision and Pattern Recognition,vol. 2 (2000) 152-159.

[14] Hafner, W., and Munkelt, O.: Using Color for Detecting Persons in Image Sequences, Pattern Recognition and Image Analysis, Vol. 7, No. 1 (1997) 47-52.

[15] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, Multimedia Mining, WSEAS Transactions on Systems, Issue 10, Volume 3, December 2004, pp. 3263-3268