GRAPH BASED DESCRIPTOR EVALUATION TOWARDS AUTOMATIC FEATURE SELECTION

Anita Keszler, Levente Kovács and Tamás Szirányi
Distributed Events Analysis Research Group, Computer and Automation Institute
{*keszler, levente.kovacs, sziranyi*}*@sztaki.hu*

Keywords: graph representation, components, feature selection

Abstract: This paper presents the first steps towards an automated image and video feature descriptor evaluation framework, based on several points of view. First, evaluation of distance distributions of images and videos for several descriptors are performed, then a graph-based representation of database contents and evaluation of the appearance of the giant component is performed. The goal is to lay the foundations for an evaluation framework where different descriptors and their combinations can be analyzed, with the goal of later application for automatic feature selection.

1 INTRODUCTION

Recent multimedia databases require large amounts of memory and computing power for storage and processing, and there is a need to efficiently index, store, and retrieve the embedded information.

Usual methods in dimensionality reduction involve various areas including Principal Component Analysis, clustering methods, neighbour searching algorithms, and so on. One of the recent approaches involves graph based clustering and component analysis methods. The generic questions and open issues here remain how to build the graphs (regarding selection of edges and weights), and how to navigate the graphs efficiently (i.e. neighbourhood searching). Numerous graph based methods have been published in the area of image/video clustering and retrieval. However, there are still several open questions in image processing and graph theory as well. Among several other problems, one of the most important step is to select the proper descriptors.

Feature selection algorithms typically can be categorized into feature ranking (based on a metric) and subset selection (optimal set of features) methods. The most common selection techniques are some sort of subset selection (e.g. wrappers). In this case the high dimensionality can be considered. In (Chapelle et al., 2002) a SVM based feature selection method is applied where the scaling factors are adjusted using the gradient of a theoretical upper bound on the error rate.

We propose and work towards practically proving that graph theoretic approaches can be useful in descriptor selection as well. As others have also begun to investigate, we build our approach on the investigation of image/video distance distributions according to several descriptors and analysing their relation and behaviour w.r.t. component formulation and giant component appearances.

In (Zhang et al., 2010) a query by example approach where histograms of point distances are investigated for 2 vs 100 feature dimensions for low number of vertices (250), as a basis to show that with increased dimensions the distance distribution of points tend to be narrower (distances decrease, showing poorer discrimination). An approach for feature selection in the presence of irrelevant features (noise) is introduced in (Sun et al., 2010), taking into consideration of sample datapoints in 2D space for decision boundary selection and investigating the distribution of feature weights in high dimensions. Another method (Morris and Kender, 2009) is based on an approx. 1000 features on real videos, using a heuristic algorithm for feature retention, and using the sort-merge approach for selecting ranked feature groups. A method for sports video feature selection is presented in (Shen et al., 2007) using Mean Shift clustering based on motion signatures and colour features. Setia et al. (Setia and Burkhardt, 2006) present a method for automatic image annotation by a feature weighting scheme and SVM using a combination of colour, texture and edge features.Guldogan et al. (Guldogan and Gabbouj, 2008) present an automatic feature selection approach for finding features

Figure 1: Sample frames from a small subset of different video contents.

that best describe a specific dataset investigating mutual information and principal component analysis.

Conversely to other approaches, in our case we do not use artificial weighting of features and distances or a priori clustering or machine learning steps, but instead use real data with multiple features and weigh the built distance graphs by the points' differences according to a specific feature, and investigate the behaviour of the difference distributions. We investigate the possible connections between distance distribution histograms and the appearance of the giant component in random geometric graphs, which are the closest representation of point vs. difference (i.e. distance) behaviour in real life datasets. The goal is to show that the method of difference distribution analysis is a good alternative to previous methods to find features with higher discrimination. We intend to lay the foundations of a framework for automatic feature evaluation and selection based solely on the descriptor difference statistics and the respective graph analysis with the goal of finding the best possible selection of descriptor combinations for the representation of related image and video contents.

2 DESCRIPTOR EVALUATIONS FOR FEATURE SELECTION

2.1 Dataset

The dataset used for the performed tests was a video database collected from real television captures, resized to 320 pixel width. The captures consist of various content categories, e.g. sport , nature , cartoons, news, street surveillance, outdoor, indoor, also containing various types of camera motions, shot lengths, and scene contents. Some example frames from various videos are shown on Fig. 1. The videos were cut into shots by our automatic shot detector, resulting in 6900 video shots with various length. The total time length of the dataset videos is 515.82 minutes.

For each shot a representative frame was extracted (based on colour histograms). When running image-based descriptors, these representative frames are used as the input for a shot. When running shot/video-based descriptors, the entire shot is used as an input. For this dataset, we extracted all the features for images and video segments, and we also calculated the distances of each element from all the others (thus enabling the creating of fully connected distance graphs based on the extracted features).

2.2 Descriptors and distance measures

With the intent of evaluating various features for general distribution and content differentiation, we selected a set of descriptors. Some are standard MPEG-7 descriptors (Manjunath et al., 2001), but we also use other features as well: local binary patterns (LBP) (Ojala and Pietikainen, 2002), curvelets (Candes et al., 2006), colour segmentation (Mean-Shift (Comaniciu and Meer, 2002) based). Further features were developed by us (average colour, relative focus regions (Kovács and Szirányi, 2007), average motion).

For calculating the differences between images/videos, we need to take into consideration which types of information the extracted features contain. For each descriptor, we used an Euclidean distance metric, i.e. for a feature all elements can be displayed along a 1D axis from 0 to $d_{max}(D)$ (maximal difference for the descriptor) and they all adhere to the triangle inequality.

In the following we list the feature contents and the used distance calculations for the used descriptors:

- MPEG-7 features: see (Manjunath et al., 2001) for a comprehensive description of the extracted feature and the used metrics.
- LBP and curvelets: Euclidean distance.
- Average colour: calculates the average colour for image blocks and produces a quantized histogram of such colours. Colour segmentation: calculates an image where different region classes are colour coded and a quantized histogram is produced; the difference between two quantized histograms in both cases is calculated as the sum of absolute differences (SAD) of the 2 normalized histograms: $d(h_1,h_2) = SAD(h_{1,N}, h2_{1,N}) = \sum_{i=0}^{4096} |h_{1,N}(i) - h_{2,N}(i)|$ where

$$h_{1,N}(i) = \frac{h_1(i) - h_{1,min}}{h_{1,max} - h_{1,min}} \qquad (1)$$

  and similarly for $h_{2,N}$.
- Relative focus regions: blurred/focused region extraction based on (Kovács and Szirányi, 2007); produces a relative focus map; the difference between two focus maps is the sum of squared differences: $SSD(f_1, f_2) = \frac{1}{fw*fh} \cdot \sum_{i=0}^{fw*fh} (f_1(i) -$

$f_2(i))^2$ where fw, fh are the width and height of the maps.

- Average motion: calculates the average motion direction for frame blocks and produces a direction histogram for the video segment on which it was run; the differences are calculated as the root squared difference of the two histograms: $d(h_1, h_2 = \sqrt{\sum_i (h_{1,i} - h_{2,i})^2}$.

3 Distributions of pairwise distances

The graph of elements is built as follows. The vertex set of the graph models the images/videos, the distance between them is calculated using the extracted features. The edge weights of the graph are proportional to the distances. The distribution of the resulting distance values is then analysed. The test results show that depending on the descriptor used to calculate the distances, the distance distributions can differ in important aspects, which we intend to exploit.

The investigation of distance distributions among database elements according to different descriptors provides information about the discriminative properties of a certain descriptor (Fig. 2 shows some examples of such distributions). As others (Zhang et al., 2010) have shown for small point sets, difference distribution behaviour can be a basis for descriptor selection (or dropping). In our case, we produced distance distributions for approx. 7000 elements from the database, applying 13 different descriptors (calculating the distance of each image/video from each other, for all descriptors). Our empirical results show that descriptors which produce distance histograms with the main peak near 0 will be less discriminative than others (i.e. most of the elements gather in one group).

4 Giant components and phase transition in random networks

The appearance of the giant component is a well known phenomenon and it was investigated in several papers, but mostly in random networks. The results known in this topic correspond to theoretical results on the existence of the giant component, and measurements on the exact threshold where the giant component first appears. Applications are usually restricted to the ER-model, however in recent years random geometric graphs (RGG) have received more attention.

These type of graphs have the ability to model networks, where the edge weights are not independent, for example derived from a metric between the vertices, and this model stands closer to modelling real image/video datasets. In this section we will give a short overview on the problem of the giant component, the random graph models, and the known results.

4.1 ER-model

Erdos and Renyi analysed the properties of a random graph with uniformly distributed edges (Erdos and Renyi, 1960). They considered the *evolution* of the components, while adding randomly selected edges to the graph. The process starts with n vertices and 0 edges, and in each step a randomly selected new edge is added independently of the already chosen edges. Recent results connected to this problem are formulated using the number of vertices, and the p probability of the existence of an edge ($G(n,p)$). p is usually described as a function of a parameter c: $p = c/n$.

Part of the theorem of Erdos and Renyi presented in (Erdos and Renyi, 1960) can be formulated as follows:

Theorem 1. *(Erdos-Renyi) The behaviour of the ER-graph from the point of view of component sizes can be divided into three important phases: The size of the largest component is denoted by C_{max}*

1) $c < 1$: $C_{max} = O(lnn)$. (The graph has small components)

2) $c = 1$: $C_{max} = \Theta(n^{2/3})$.

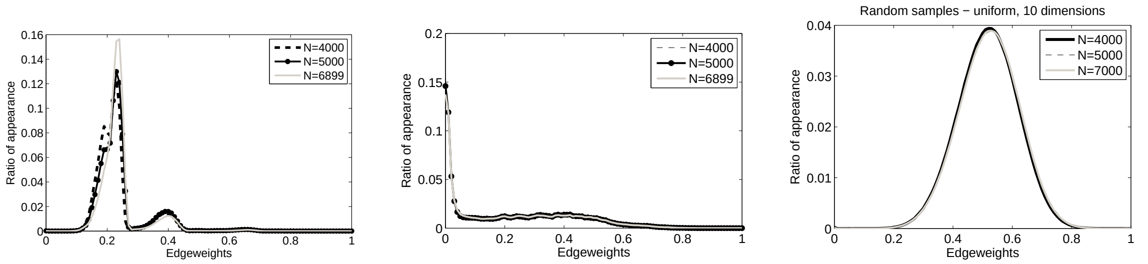
3) $c > 1$: $C_{max} = O(n)$ (giant component), but all the other components have size $O(lnn)$.

The results presented in (Erdos and Renyi, 1960) also deal with the complexity of the components, but now we only interested in their size.

4.2 Random geometric graphs

Besides the above mentioned classic random graph models, several different versions have been published in order to model certain properties of complex real networks. For example in case of distance graphs, the edge-weights corresponds to pairwise distances of objects based on a given metric. Random geometric graph models offer a solution to mimic these type of dependencies, since in this model, the edge-weights are not selected independently of each other (Penrose, 2003).

Definition 1. *A **random geometric graph** or RGG is obtained as follows. We take $X_1, X_2, \ldots, X_n \in \mathbb{R}$ at random (according to some probability distribution ν on $\mathbb{R}^d$, where d is the number of dimensions). For $i \neq j$ we connect X_i and X_j if $\| X_i - X_j \| < r_n$. r_n is the radius of the random geometric graph.*



(a) Average colour based distance distributions (b) Focus histogram based distance distributions (c) Random samples based distance distributions

Figure 2: Distance distribution histograms for different descriptors.

The existence of the giant component in random geometric graphs has also been examined. *Thermodinamic limit*, a term of statistical physics was used to describe this phenomenon. This limit corresponds to the critical radius of the RGG: $r_n \sim c \cdot n^{-1/d}$. At this limit, the expected value of the average degree in the graph tends to a constant. Above a certain c constant in the formula of the radius r_n^d there is likely to be a giant component.

Unfortunately, the exact value of r_n is unknown. The fact, that $0 < r_n < \infty$, if $d \geq 2$ is an interesting result of continuum percolation, in itself.

Our aim with the tests on random geometric graphs is to analyse the correspondence between the critical value r_n (or in other words the critical edge weight), the dimension of the RGG and the number of vertices. Although we cannot give exact values, the tendencies are also important in real applications, and for our purposes.

In Fig. 3 the correspondence between the previously mentioned parameters is shown. It is important to note, that the number of dimensions has a significant impact on the critical edge weight in case of RGGs. There is a difference between the results in case of different graph sizes as well, but it is not that relevant compared to the number of dimensions.

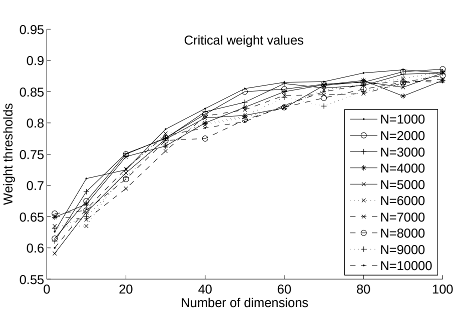


Figure 3: Critical weight values of artificial datasets.

5 Giant components in real datasets

The appearance of the giant component in real networks with geometric restrictions on the edge weights is an interesting mathematical topic on its own, but it also has the potential to be used in applications where the structure of the evolving graph is important as well.

In retrieval and clustering tasks finding the 'optimal' graph structure is still an open question. Algorithms for weighted graphs are usually more complex than the ones for non-weighted graphs. Due to this fact, several applications transform the originally weighted graph into a non-weighted one for processing. In this case the question arises: how to transform the weighted graph? The generally applied solution is to select a distance threshold; if the distance between two vertices is lower than the threshold, the vertex pair will be connected. Studying the graph structure with different thresholds is a key step for selecting this threshold.

On one hand, our tests intent to analyse the evolving components at different thresholds for a given descriptor and the corresponding metric. On the other hand, graphs of different descriptors are compared, and based on the these test results, we propose a new aspect of comparing the descriptors themselves.

5.1 Appearance of the giant component

The 'descriptor graphs' were analysed to find out whether the appearance of the giant component is traceable, and if it is, how does it depend on the selected descriptor. By definition, the giant component is a component with size $O(n)$, while all other components have size $O(logn)$, so the exact critical threshold can be determined by the analysis of the asymptotic behaviour of the network. In the case of our tests on real data this type of test can not be carried out, since we are working with a finite number of vertices, but acceptable estimations are available.

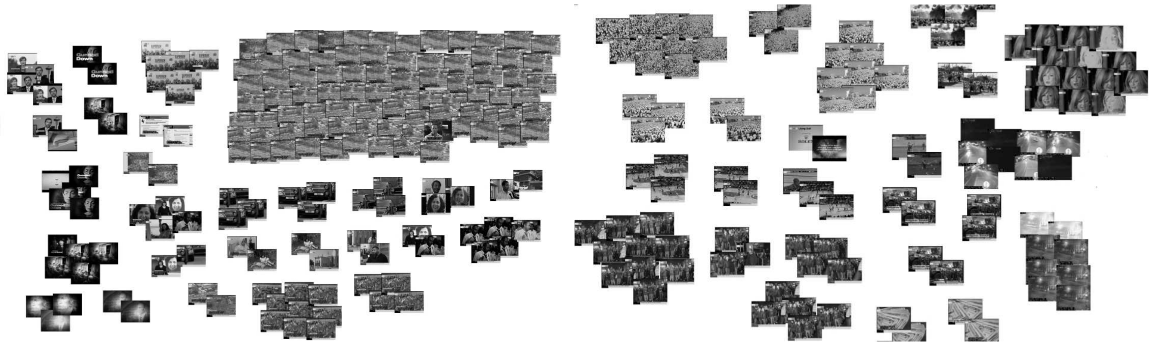


Figure 4: Example of components at an inner building step of the Edge histogram graph. The similar images tend to correspond in the same component.

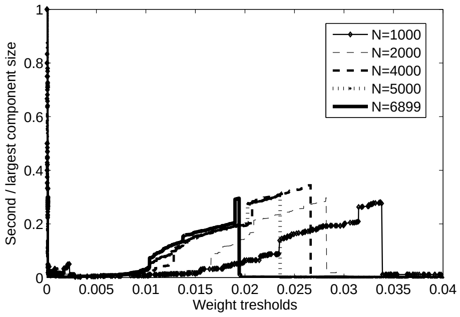


Figure 5: Ratio of the second and the largest component sizes of the Focus histogram graph.

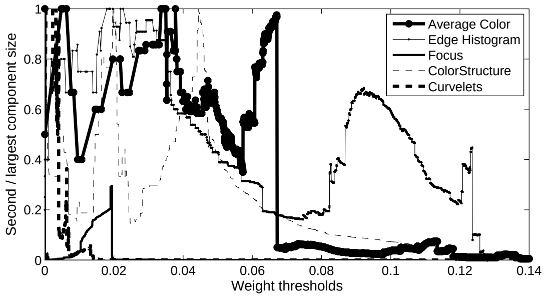


Figure 6: Ratio of the second and the largest component in case of different descriptors.

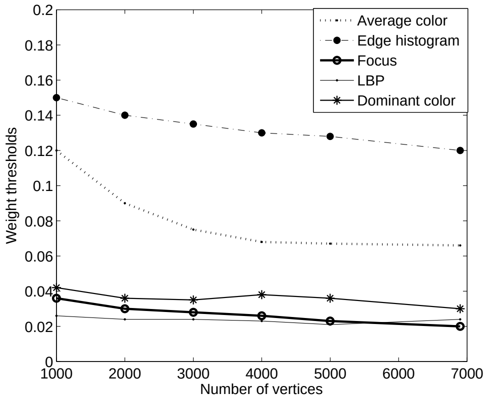


Figure 7: Critical weight values in the graph of different descriptors.

The stopping condition for selecting edges in the descriptor graph is the estimated threshold, where the giant component appears. In our finite graphs the ratio of the second and the largest component will be the parameter to track the evolving components.

5.2 Ratios of largest and second largest components

The ratio of the largest and second largest component is relevant with relation to descriptor behaviour, during the building of the graph in the search for the point where the single giant component appears. This ratio shows different behaviour w.r.t. simulated datasets, showing multiple peak regions during the process. This means, that during the graph edge/weight selection there are intervals when multiple components grow in parallel, which is inline with the expectations that when multiple content classes of images/videos are in the same dataset, a descriptor will produce denser regions, containing smaller components with similar contents.

As it was mentioned, the ratio of the largest and second largest component is important in relation to descriptor behaviour, during the building of the graph, in the search for the point where the single giant component appears. This ratio (examples shown in Fig.5 and 6) shows different behaviour w.r.t. simulated datasets, with multiple peak regions during the process. This solidifies the expectations that different components growing in parallel will produce separate denser regions containing inter-similar contents.

The critical edge weights of some descriptor graphs are shown on Fig. 7. As shown, the critical weights depend on the number of vertices of the graph, but the impact of this parameter depends on the descriptor. Detailed test results on the critical weight value of the Focus descriptor graph (see Fig.5) shows how it depends on the size of the graph.

Depending on the task, the estimated values of the critical edge weights have high importance. For example in shortest-paths based clustering task (label propagation) it is necessary to know whether labels could spread through the graph - i.e. is the graph connected (or almost connected). In this case the optimal threshold would be near the phase transition's critical value. On the other hand, if we are interested in the selection of possible cluster cores (dense regions), we should select a threshold that results in a graph with small dense components.

Our work presents the possibility of finding the optimal threshold, depending on the selected descriptor. This way we are also able to evaluate the 'quality' of a descriptor. The lower the critical weight value is, the smaller the chance of finding relevant dense cluster cores.

6 CONCLUSIONS AND FUTURE WORK

This paper presents the first steps towards an automatic feature selection framework, investigating descriptor behaviour based on the analysis of random geometric graphs structures built from real data and by using element distances based on several descriptors and their distance / difference distributions along with the generic behaviour of such graph types during the appearance of the giant component. Our next goal is to produce an descriptor evaluation framework which analyses graph-connectedness weighted by difference distributions and their relation to the thresholds associated to the estimated appearances of the giant component, and rank descriptors (and combinations of descriptors) based on these properties. Histograms of such distances combined with graph analysis based on random graph theory can provide a solid foundation for image and video feature selection.

ACKNOWLEDGEMENTS

This work has been partially supported by Hungarian Scientific Research Fund grants 83438 and 80352.

REFERENCES

Candes, E., Demanet, L., Donoho, D., and Ying, L. (2006). Fast discrete curvelet transforms. *Multiscale Modeling and Simulation*, 5(3):861–899.

Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159.

Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(5):603–619.

Erdos, P. and Renyi, A. (1960). On the evolution of random graphs. Pbulication of the Mathematical Institute of the Hungarian Academy of Sciences.

Guldogan, E. and Gabbouj, M. (2008). Feature selection for content-based image retrieval. *Signal, Image and Video Processing*, 2(3):241–250.

Kovács, L. and Szirányi, T. (2007). Focus area extraction by blind deconvolution for defining regions of interest. *IEEE Tr. on Pattern Analysis and Machine Intelligence*, 29(6):1080–1085.

Manjunath, B. S., Ohm, J. R., Vasudevan, V. V., and Yamada, A. (2001). Color and texture descriptors. *IEEE Trans. on Circuits and Systems for Video Technology*, 2(6):703–715.

Morris, M. and Kender, J. (2009). Sort-merge feature selection and fusion methods for classification of unstructured video. In *Proc. of IEEE international conference on Multimedia and Expo*, pages 578–581.

Ojala, T. and Pietikainen, M. (2002). Multiresolution grayscale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on attern Analysis and Machine Intelligence*, 24(7).

Penrose, M. (2003). *Random Geometric Graphs*. Oxford University Press.

Setia, L. and Burkhardt, H. (2006). Feature selection for automatic image annotation. In *Proc. of 28th Pattern Recognition Symposium of the German Association for Pattern Recognition*. Springer.

Shen, Y., Lu, H., and Xue, X. (2007). A semi-automatic feature selecting method for sports video highlight annotation. In *Proc. of 9th Intl. Conference on Advances in visual information systems*, pages 38–48.

Sun, Y., Todorovic, S., and Goodison, S. (2010). Local learning based feature selection for high dimensional data analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1610–1626.

Zhang, W., Men, S., Xu, L., and Xu, B. (2010). Feature distribution based quick image retrieval. In *Proc. of Web Information Systems and Applications Conference*, pages 23–28.