SEARCH AND RETRIEVAL OF MULTI-MODAL DATA ASSOCIATED WITH IMAGE-PARTS

Niloufar Pourian, S. Karthikeyan, B.S. Manjunath

Department of Electrical and Computer Engineering, University of California, Santa Barbara, USA
{ npourian, karthikeyan, manj } @ ece.ucsb.edu

ABSTRACT

We present a novel framework for querying multi-modal data from a heterogeneous database containing images, textual tags, and GPS coordinates. We construct a bi-layer graph structure using localized image-parts and associated GPS locations and textual tags from the database. The first layer graphs capture similar data points from a single modality using a spectral clustering algorithm. The second layer of our multi-modal network allows one to integrate the relationships between clusters of different modalities. The proposed network model enables us to use flexible multi-modal queries on the database.

***Index Terms*—** Search and Retrieval, Multi-Modal, Community Detection

1. INTRODUCTION

In recent years, the amount of information with various modalities has rapidly increased. This has been more noticeable in digital image collections. Every day, millions of images associated with noisy multi-modal labels are generated. In many scenarios in social media mining, medical image analysis and surveillance, analysts benefit from representing queries using multiple information sources. For example, radiologists interpreting Traumatic Brain Injury (TBI) have access to MRI, CT imaging data and patient attributes such as age, location, and ethnicity. It would be of a significant help if they can query such a heterogeneous database using a combination of these features, such as extent of the injury for a patient age group of 16-20 with a specific ethnicity. As another example, in a surveillance scenario, the analyst might be interested in determining geo-localized instances of a specific car within a specific time-interval. To address such "multi-modal" queries new methods need to be developed that effectively integrate information derived from heterogeneous datasets.

Most of the multimedia retrieval techniques focus on the retrieval of single modality data, such as images [1, 2, 3] and text documents [4, 5]. Some recent works have focused on cross-media retrieval where the query examples and the query results have different modalities [6, 7, 8, 9]. The main difficulty in cross-media retrieval is to define a similarity measure among heterogeneous low-level features.

In order to simultaneously search and retrieve data from multiple modalities, other approaches have been considered [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. For instance in [16], it is experimentally shown that multimodal queries achieve higher retrieval accuracy than mono-modal ones. The work in [10] suggests using a combination of ontology browsing and keyword-based querying. The methods presented in [11, 14, 15, 16] use a similar approach and rely on the assumption that every document has an equal number of nearest neighbors for each of the modalities. However, such an assumption might degrade the retrieval performance as a document containing "image+text" may have many nearest neighbors in image modality, but not as many relevant textual data. Finally, all these approaches are generally based on a global image understanding rather than a localized one. Therefore, they are unable to provide the most relevant multi-modal data for different image-parts.

This paper focuses on information retrieval associated with image-parts. Understanding the relevance between an image and data points of different modalities such as textual tags or GPS locations strongly depends on how well the image itself is understood. The proposed formulation introduces a way of understanding different image-parts and associates each image-part with its most relevant multi-modal information. Our approach utilizes the co-occurrence of data points within a modality and across multiple modalities using a graph structure. We note that the proposed work does not impose relevance between each document and an equal number of media objects from all modalities as assumed by [11, 14, 15, 16]. In summary, our contributions are threefold:

- We introduce a retrieval system by constructing mono-modal graph structures and detecting communities of related data points from a single modality. In particular, a graphical image structure enables the system to learn different image-parts.
- By learning coherent communities among graph structures of different modalities, we create a network to integrate the relationships between communities.
- Relevant multi-modal data for a given image-part is retrieved using a hierarchical search.

This work was supported by ONR grant N00014-12-1-0503.

2. APPROACH

In this paper, different types of metadata are treated as different modalities. For illustration purposes and without loss of generality, we focus on three different modalities: 2D images, textual tags, and GPS locations. However, proposed method can be generalized to an arbitrary number of modalities.

2.1. Graph Structure: Single Modality

We incorporate the similarity between the data points of a single modality m by creating a graph structure $G^{(m)} = (V^{(m)}, A^{(m)})$ with $V^{(m)}$ representing the nodes of graph $G^{(m)}$ from modality m, and $A^{(m)}$ being the corresponding adjacency matrix. In what follows, we describe how each of these graph structures are created in details.

Modality 1 - Image: The purpose of creating a graphical image structure is to group similar objects together. Following the work of [21, 22], we integrate the visual characteristics along with the spatial information of image-parts across all database images in a graph structure. To provide spatial information, we utilize a segmentation algorithm based on color and texture [23]. The graphical image structure $G^{(1)} = (V^{(1)}, A^{(1)})$ contains $\sum_{I=1}^{D} |v_I|$ number of nodes, with $|v_I|$ denoting the number of segmented regions of image I, and D representing the total number of images in the database. Two nodes i and j are connected if they are spatially adjacent or if they are visually similar:

$$A_{ij}^{(1)} = \mathcal{I}(i \in \mathcal{F}_j \quad or \quad i \in \mathcal{H}_j), \quad \forall i, j \in V^{(1)} \tag{1}$$

where $\mathcal{I}(x)$ is equal to 1 if x holds true and zero otherwise. In addition, $\mathcal{F}_j$ indicates the set of all nodes (segmented regions) that are visually similar to node j and $\mathcal{H}_j$ is the set of all nodes in the spatial neighborhood of node j. To represent each image, DenseSift features [24] are extracted and then quantized into a visual codebook. To form a regional signature h^i for a node i, features are mapped to the segmented regions that they belong to. Then, a histogram representation of the codebook is created. Two nodes i and j are considered visually similar if the normalized distance between the two nodes' visual features is less than a threshold $\alpha_I \in \mathbb{R}^+$. The distance between two regional histograms is measured by the Hellinger distance, as it has been shown to be a good metric for computing the distance between histograms in retrieval problems [25].

Modality 2 - Textual tags: We use graph $G^{(2)} = (V^{(2)}, A^{(2)})$ to encode the co-appearance of textual tags. Two nodes (textual tags) are connected if they are highly related. The adjacency matrix for this graph structure $A^{(2)}$ is defined by:

$$A_{ij}^{(2)} = \mathcal{I}(i \in \mathcal{T}_j), \quad \forall i, j \in V^{(2)} \tag{2}$$

with $\mathcal{T}_j$ denoting the set of all nodes that have a co-appearance distance of less than $\alpha_t \in \mathbb{R}^+$ from node j. To measure the distance between two textual tags i and j, binary vectors t_i and t_j are defined to capture the presence or absence of each tag across the database. The distance between the two binary tag vectors is computed using the Cosine distance [5].

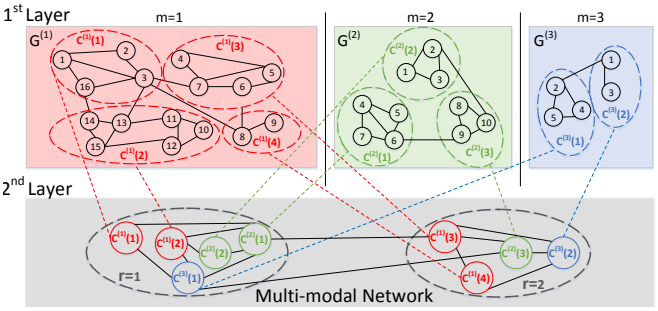

Fig. 1: Illustration of the graphical structures in the first layer and the multi-modal network representation in the second layer. Figure is best viewed in color.

Modality 3 - GPS coordinates: Graph $G^{(3)} = (V^{(3)}, A^{(3)})$ illustrates the GPS data locations in the dataset. Two nodes are connected if and only if (iff) their locations have a close distance. Here, $A^{(3)}$ is defined by:

$$A_{ij}^{(3)} = \mathcal{I}(i \in \mathcal{G}_j), \quad \forall i, j \in V^{(3)} \tag{3}$$

with $\mathcal{G}_i$ indicating the set of all nodes (GPS locations) close to node i. Two GPS locations i and j are considered to be close if their Euclidean distance is less than a threshold $\alpha_g \in \mathbb{R}^+$.

2.2. Community Detection: Single Modality

Our goal is to find groups of similar/related single modality data in each of the graph representations. This is done by applying a graph partitioning algorithm to each of the graphical structures. Each of these groups resembles a bag of related single modality data and is referred to as a "community".

To perform the graph partitioning, we use a normalized cut method as described in [26]. In this algorithm, the quality of the partition (cut) is measured by the density of the weighted links inside communities as compared to the weighted links between communities. The objective is to maximize sum of the weighted links within each community while minimizing sum of the weighted links across community pairs.

Let M be the number of modalities. A graph $G^{(m)}$ with $m \in \{1, \ldots, M\}$ consists of $C^{(m)}$ detected communities. Each detected community $c^{(m)}(\nu)$ with $\nu \in \{1, \ldots, C^{(m)}\}$ is a collection of related nodes in that particular modality. For example, in the graphical image structure, each community contains all the pieces/image-parts of an object and mapping these back to each segmented image would highlight/detect that particular object [21, 22]. A detected community for the textual tags' graph corresponds to highly related/correlated tags. Finally, a community for the GPS graph contains all the GPS locations that are relatively close to each other.

Community Representation: Let $\mathcal{P}^{(m)}(\nu)$ denote the set of all nodes in community $c^{(m)}(\nu)$, with $\nu \in \{1, \ldots, C^{(m)}\}$

and $m \in \{1, \ldots, M\}$. Each detected community in a graphical image structure is represented by the average of the histograms representations of the nodes it contains. This is summarized by $c^{(1)}(\nu) = \sum_{p \in \mathcal{P}^{(1)}(\nu)} h_p / |\mathcal{P}^{(1)}(\nu)|$. Community $c^{(2)}(\nu)$ in the textual-tag graph structure is represented by a binary vector of length K_t with each bin k being equal to 1 iff the corresponding tag exists within that cluster. This can be summarized by $c_k^{(2)}(\nu) = \mathcal{I}(\sum_{p \in \mathcal{P}^{(2)}(\nu)} t_{pk} > 0) \quad , k = \{1, \ldots K_t\}$. Finally, each community $c^{(3)}(\nu)$ is represented by the average of all the nodes (GPS coordinates) that it contains: $c^{(3)}(\nu) = \sum_{p \in \mathcal{P}^{(3)}(\nu)} g_p / |\mathcal{P}^{(3)}(\nu)|$.

2.3. Network Representation of Multi-Modal Data

A second layer multi-modal network is constructed to integrate the relation among the multi-modal data. This network is denoted by $W = (V, A)$ with V and A representing the nodes and the adjacency matrix of W, respectively. Each node in this network is a detected community in a mono-modal graph structure, and itself is a collection of nodes in the first layer (mono-modal data points). This can be summarized by: $V = \bigcup_{m=1}^{M} \bigcup_{\nu=1}^{C^{(m)}} c^{(m)}(\nu)$, where M is the total number of modalities.

In this network, node $c^{(m)}(\nu)$ from modality m and node $c^{(m')}(\nu')$ from modality m' are connected if the co-occurrence rate of the nodes fallen into community $c^{(m)}(\nu)$ in the first layer with the nodes fallen into community $c^{(m')}(\nu')$ in the first layer is larger than a threshold γ. Thus, adjacency matrix A of the second layer can be summarized as:

$$A_{c^{(m)}(\nu), c^{(m')}(\nu')} = \mathcal{I}\left(\frac{\sum_{i \in c^{(m)}(\nu), j \in c^{(m')}(\nu')} \mathcal{E}_{i,j}^{m,m'}}{\sum_{i \in G^{(m)}, j \in G^{(m')}} \mathcal{E}_{i,j}^{m,m'}} \geq \gamma \right) \quad (4)$$

We set $\mathcal{E}_{i,j}^{m,m'} = 1$ if node node i from modality m co-occurs with node j from modality m' and zero otherwise. To make a fair comparison among different modalities, we normalize each co-occurrence rate by the maximum co-occurrence rate that can occur between the two particular modalities.

2.4. Multi-modal Community Detection

To group highly similar/related nodes across different modalities (second layer), we use a graph partitioning algorithm on the multi-modal network. It is worth noting that the first layer of clustering allows one to combine the similar single modality data points together. The second layer of clustering is introduced to integrate/learn highly related data across different modalities. Our approach is favorable as it does not depend on bringing multi-modal data into a common feature space. In addition, the proposed approach integrates the information from each of the modalities independent of its other co-existing modalities. This implies that if a collection of images are accompanied by textual tags and GPS locations and a similar image is accompanied by a similar textual tag but no GPS information, our model is expected to group such multi-modal information together and consequently associate the image with the corresponding GPS location.

In the remainder of this paper, let R denote the total number of communities in the second layer network and $r \in \{1, \ldots, R\}$ represent each of the detected multi-modal communities.

3. QUERY RETRIEVAL

When a multi-modal query q is presented to the system, the most relevant multi-modal data are retrieved and ranked based on three different factors: the strength of relevance of each multi-modal cluster to the query, the strength of relevance of each mono-modal cluster within that multi-modal cluster, and finally the degree of similarity between each data point in each mono-modal cluster and the query.

The level of relevance between each multi-modal cluster r and a query q is measured by computing the similarity between the query data and the mono-modal clusters in r. This is summarized by:

$$\Lambda_r = \frac{\sum_{m \in M_q} \sum_{u=1}^{U_r^{(m)}} \sum_{x=1}^{Q^{(m)}} e^{-d(c^{(m)}(u), q^{(m)}(x))}}{|M_q| \times U_r^{(m)} \times Q^{(m)}}, \quad (5)$$

with $r \in \{1, \ldots, R\}$, M_q representing the set of all modalities present in the query, $U_r^{(m)}$ denoting the total number of mono-modal clusters of modality m within the multi-modal cluster r, and $Q^{(m)}$ representing the total number of data points from modality m within the query. In addition, $q^{(m)}(i)$ denotes the ith query data from modality m and $d(.,.)$ is the Euclidean distance function. A multi-modal cluster with a higher similarity score has the highest priority score.

The strength of relevance of each mono-modal cluster within the multi-modal cluster r is measured by taking into account the amount of similarity between the query and the mono-modal cluster, the similarity of the query with other mono-modal clusters in r, as well as the density of the links between that cluster and other clusters. The strength of relevance of each mono-modal cluster $c^{(m)}(\nu)$ is given by:

$$\Omega_{c^{(m)}(\nu)} = \frac{1}{Q^{(m)} U_r^{(m')} Q^{(m')}} \sum_{x=1}^{Q^{(m)}} e^{-d(q^{(m)}(x), c^{(m)}(\nu))} \times \sum_{u=1}^{U_r^{(m')}} \sum_{y=1}^{Q^{(m')}} f(c^{(m)}(\nu), c^{(m')}(u)) \; . \; e^{-d(q^{(m')}(y), c^{(m')}(u))}, \quad (6)$$

with $\nu \in \left\{1, \ldots, U_r^{(m)}\right\}$, and m' representing all modalities except modality m. The function $f(c^{(m)}(\nu), c^{(m')}(u))$ denotes the normalized density of the links between clusters $c^{(m)}(\nu)$ and $c^{(m')}(u)$ and is defined by the total number of links between cluster $c^{(m)}(\nu)$ and cluster $c^{(m')}(u)$ divided by the maximum number of links between cluster $c^{(m)}(\nu)$ and any other cluster from modality m'. This is followed by normalizing each $\Omega_{c^{(m)}(\nu)}$ by the maximum possible value of $\Omega_{c^{(m)}(\nu)}$ for all existing modalities. Single modality data points within a mono-modal cluster $c^{(m)}(\nu)$ with larger values of $\Omega_{c^{(m)}(\nu)}$ are ranked higher.

Finally, the degree of similarity between a data point $i^{(m)}$ in mono-modal cluster $c^{(m)}(\nu)$ and the query q is computed based on the similarity between $i^{(m)}$ and its matching modality data given in q as well as the similarity between data points from modality m' within the multi-modal cluster r that are connected to $i^{(m)}$ and the query. This formulation allows one to not only include the similarity of q with point $i^{(m)}$, but also the relativity of connected nodes to $i^{(m)}$ with q. We give a higher priority to data points from modalities that exist within the query q. Data points from modalities not present in q are ranked based on their strength of connectivity with data points from modalities present within q. Such a ranking process is summarized in Figure 3.

4. EXPERIMENTAL RESULTS

Datasets: The experiments are conducted on two challenging datasets: VOC07, and MMFlickr. The VOC07 dataset is publicly available and contains 2 different modalities: $9,963$ 2D images and 20 textual tags. We created the MMFlickr dataset to simulate scenarios in real world where data object from modality m is not necessarily accompanied by data object from all other modalities. This dataset contains three different modalities of 2D images, textual tags, and GPS information and is consisted of $1,320$ data objects across three different modalities. MMFlickr was created by crawling data from Flickr photo sharing web page using the Flickr API [27].

Performance: In our experiments, a set of multi-modal queries are defined and the accuracy of the retrieval system is measured using f-measure [28]. The performance is compared with the state-of-the-art multi-modal retrieval system presented in [16].

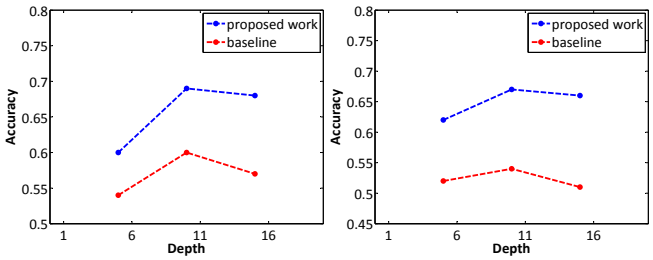

Fig. 2: Comparison of the performance of the proposed algorithm with the baseline approach. Results are reported for VOC07 (left) and MMFlickr (right) databases.

As shown in Figure 2, our approach is able to achieve a higher retrieval accuracy than the baseline method without imposing the unnecessary constraint that every document has an equal number of nearest neighbors for each of the modalities. It is also interesting to note that the accuracy gap between the proposed work and the baseline approach is slightly larger at higher depths for both VOC07 and MMFlickr datasets. These results emphasizes the applicability of the proposed method.

Figure 4 illustrates some sample queries along with their corresponding results of the proposed multi-modal retrieval system. As shown, our method is able to return multi-modal data related to the presented queries.

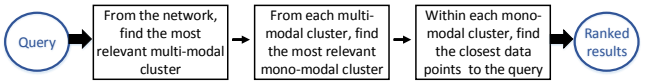

Fig. 3: Ranking process for a given query as descibed in Section 3.

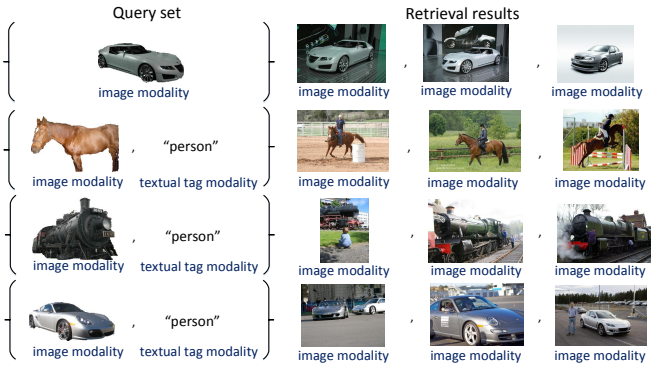

Fig. 4: Sample queries and their corresponding top 3 multi-modal retrieval results using the proposed system on VOC database. Figure is best viewed in color.

We speculate that the best retrieval results are achieved for data points that have co-occurred in the dataset often. On the other hand, if a query set is a combination of multi-modal data points that rarely co-occur, the top retrieval results only reflect those particular data points in the query set that co-occur often. Figure 5 demonstrates such query along with the top 3 retrieval results of the proposed approach. In this example, although retrieval results are relevant to some query data points, they do not correspond well to all data points in the query set. In order to see retrieval results from multiple modalities, larger number of top retrieved results needs to be shown.

Fig. 5: A sample query set with datapoints rarely co-occuring and its corresponding top 3 multi-modal retrieval results using the proposed system. Figure is best viewed in color.

5. CONCLUSION

This paper presented a new model to provide a rich set of multi-modal data for localized image-parts. We introduced mono-modal graph structures and applied a spectral clustering algorithm to find related data points of a single modality. In addition, the relationship between clusters of different modalities were integrated using a multi-modal network. Experimental results conducted on two challenging datasets demonstrated that the proposed approach compares favorably with current state of the art methods. In future, we plan to explore the applicability of the proposed approach for data compression.

6. REFERENCES

[1] E. Attalla and P. Siy, "Robust shape similarity retrieval based on contour segmentation polygonal multiresolution and elastic matching," in *Pattern Recognition*, 2005.

[2] M. Kokare, P. Biswas, and B. Chatterji, "Texture image retrieval using new rotated complex wavelet filters," in *Trans. Systems, Man, and Cybernetics*, 2005.

[3] M. Worring and T. Gevers, "Interactive retrieval of color images," in *International Journal of Image and Graphics*, 2001.

[4] David C Blair and ME Maron, "An evaluation of retrieval effectiveness for a full-text document-retrieval system," *Communications of the ACM*, 1985.

[5] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," in *Information processing & management*, 1988.

[6] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. international conference on Multimedia*, 2010.

[7] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proc. ACM international conference on Multimedia*, 2009.

[8] H. Zhang and J. Weng, "Measuring multi-modality similarities via subspace learning for cross-media retrieval," in *Advances in Multimedia Information Processing-PCM*. 2006.

[9] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *ICCV*, 2011.

[10] G. Hubert and J. Mothe, "An adaptable search engine for multimodal information retrieval," in *Journal of the American Society for Information Science and Technology*, 2009.

[11] M. Lazaridis, A. Axenopoulos, D. Rafailidis, and P. Daras, "Multimedia search and retrieval using multimodal annotation propagation and indexing techniques," in *Signal Processing: Image Communication*, 2013.

[12] S. Sabetghadam, M. Lupu, and A. Rauber, "Astera-a generic model for semantic multimodal information retrieval," *Available at http://ceur-ws.org/*.

[13] M. Bokhari and F. Hasan, "Multimodal information retrieval: Challenges and future trends," *Available at http://researchgate.net/*.

[14] D. Rafailidis, S. Manolopoulou, and P. Daras, "A unified framework for multimodal retrieval," in *Pattern Recognition*, 2013.

[15] A. Axenopoulos, P. Daras, S. Malassiotis, V. Croce, M. Lazzaro, J. Etzold, P. Grimm, A. Massari, A. Camurri, and T. Steiner, "I-search: a unified framework for multimodal search and retrieval," in *The Future Internet*. 2012.

[16] P. Daras, S. Manolopoulou, and A. Axenopoulos, "Search and retrieval of rich media objects supporting multiple multimodal queries," in *Trans. Multimedia*, 2012.

[17] Edward Chang, Kingshy Goh, Gerard Sychay, and Gang Wu, "Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines," *IEEE Transactions on Circuits and Systems for Video Technology*, 2003.

[18] Stefan Romberg, Rainer Lienhart, and Eva Hörster, "Multimodal image retrieval," *International Journal of Multimedia Information Retrieval*, 2012.

[19] Ruofei Zhang, Zhongfei Zhang, Mingjing Li, Wei-Ying Ma, and Hong-Jiang Zhang, "A probabilistic semantic model for image annotation and multimodal image retrieval," in *ICCV*, 2005.

[20] Duc-Tien Dang-Nguyen, Giulia Boato, Alessandro Moschitti, and Francesco GB De Natale, "Supervised models for multimodal image retrieval based on visual, semantic and geographic information," in *International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2012.

[21] Niloufar Pourian and B.S. Manjunath, "PixNet: A Localized Feature Representation for Classification and Visual Search," *To appear in IEEE Transactions on Multimedia*.

[22] Niloufar Pourian and B.S. Manjunath, "Retrieval of images with objects of specific size, location and spatial configuration," in *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015.

[23] Y. Deng and B.S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2001.

[24] D. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, 1999.

[25] M. S. NIKULIN, *Hellinger Distance*, Springer, 2001.

[26] J. Shi and J. Malik, "Normalized cuts and image segmentation," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2000.

[27] "Flickr-Open-API," *http://www.flickr.com/services/api/*.

[28] David Martin Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.