Open Research Online

The Open University's repository of research publications
and other research outputs

# Combining visual and textual systems within the context of user feedback

## Conference or Workshop Item

How to cite:

Kaliciak, Leszek; Song, Dawei; Wiratunga, Nirmalie and Pan, Jeff (2013). Combining visual and textual systems within the context of user feedback. In: 19th International Conference on MultiMedia Modeling (MMM2013), 5-7 Jan 2013, Huangshan, China, pp. 445–455.

For guidance on citations see FAQs.

© 2013 Springer-Verlag

Version: Accepted Manuscript

Link(s) to article on publisher's website:
http://dx.doi.org/doi:10.1007/978-3-642-35725-1$_4$1
http://mmm2013.org/apl.htm

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.

oro.open.ac.uk

# Combining Visual and Textual Systems within the Context of User Feedback

Leszek Kaliciak[1], Dawei Song[2], Nirmalie Wiratunga[1], and Jeff Pan[3]

[1] The Robert Gordon University, Aberdeen, UK
[2] The Open University, Milton Keynes, UK
[3] Aberdeen University, Aberdeen, UK
{l.kaliciak,n.wiratunga}@rgu.ac.uk, Dawei.Song@open.ac.uk,
jeff.z.pan@abdn.ac.uk

**Abstract.** It has been proven experimentally, that a combination of textual and visual representations can improve the retrieval performance ([20], [23]). It is due to the fact, that the textual and visual feature spaces often represent complementary yet correlated aspects of the same image, thus forming a composite system.

In this paper, we present a model for the combination of visual and textual sub-systems within the user feedback context. The model was inspired by the measurement utilized in quantum mechanics (QM) and the tensor product of co-occurrence (density) matrices, which represents a density matrix of the composite system in QM. It provides a sound and natural framework to seamlessly integrate multiple feature spaces by considering them as a composite system, as well as a new way of measuring the relevance of an image with respect to a context. The proposed approach takes into account both intra (via co-occurrence matrices) and inter (via tensor operator) relationships between features' dimensions. It is also computationally cheap and scalable to large data collections. We test our approach on ImageCLEF2007photo data collection and present interesting findings.

**Keywords:** Visual and Textual Systems' Combination, Visual Features, Textual Features, User Feedback, Tensor Product, Density Matrix, Expectation Value.

## 1 Introduction

It has been proven experimentally (i.e. the annual imageCLEF competition results) that a combination of textual and visual representations can improve the retrieval performance ([20], [23]). It is due to the fact, that the textual and visual feature spaces often represent complementary yet correlated aspects of the same image, thus forming a composite system. This, in turn, presents an opportunity to utilize this complemetarity by combining the systems in order to improve their performance.

Visual and textual systems can be combined within the context of image retrieval or automatic image annotation. The latter exploits the relationships between the features' dimensions to automatically annotate images that do not have textual descriptions. However, even after auto-annotating the images, the retrieval system often (apart from some projection based methods, i.e. LSI) needs to combine the features in a meaningful way in order to utilize the complementarity of the aforementioned feature spaces

S. Li et al. (Eds.): MMM 2013, Part I, LNCS 7732, pp. 445–455, 2013.
© Springer-Verlag Berlin Heidelberg 2013

to improve the retrieval. Some of these combination methods can be modified to incorporate the user feedback.

This paper focuses on the combination of the systems within the context of image retrieval, and to be more precise - the context of a user feedback. The data collection that we conduct our experiments on, ImageCLEF2007photo, is a fully annotated one (albeit the description field which was present in the ImageCLEF2006 collection is now unavailable).

Thus, most approaches that combine visual and textual features in content based image retrieval systems are fusion methods that would:

1. pre-filter the data collection by visual content and then re-rank the top images by text ([4]);
2. pre-filter the data collection by text and then re-rank the top images by visual content ([5]);
3. pre-filter the data collection by visual (textual) content and then aggregate the scores of the textual (visual) representations of the top retrieved images (transmedia pseudo-relevance mechanism [6]);
4. fuse the representations (early fusion [7]);
5. fuse the scores or ranks (late fusion [8]).

This paper is organized as follows: Section 2 presents work related to the combination of visual and textual features in general. Section 3 describes the theoretical model for combination of visual and textual systems in the context of user feedback. The experimental setup and results with their discussion forms the next section, Section 4. Finally, Sections 5 and 6 are devoted to conclusions and future work, respectively.

## 2 Related Work

In this work, we modify the existing models (that combine visual and textual features) in order to incorporate user feedback. Thus modified approaches will serve as our comparison baselines.

Pre-filtering by text and re-ranking by visual content is usually a well performing method. However, the main drawback of this approach is that the images without the textual description will never be returned by the system (although one could try to auto-annotate the collection beforehand). Moreover, this type of pre-filtering relies heavily on the textual features and the assumption that the images are correctly annotated.

The most common early fusion technique is concatenation of visual and textual representations. Some recently proposed models incorporate the tensor product to combine the systems [9]. The aforementioned tensor product presents a sound fusion technique as it takes into account all of the combinations of different features' dimensions. The main drawback of the early fusion approach, however, is the well known curse of dimensionality. Later in the paper we show, that the curse of dimensionality can often be avoided as the similarity between the fused representations may be characterized as the combinations of similarities computed on individual feature spaces.

In case of the late fusion, the most widely used method is the arithmetic mean of the scores, their sum (referred to as CombSUM), or their weighted linear combination. One

of the best performing systems on the ImageCLEF2007 data collection, XRCE [10], utilizes both (for comparison purposes) early (concatenation of features) and late (an average of scores) fusion approaches. Another common combination method, referred to as CombPROD in the literature, is the square of the geometric mean of the scores - their product. It has been argued, that the major drawback of the late fusion approaches is their inability to capture the correlation between different modalities [11]. However, later in the paper we show, that in some cases the late fusion can be represented as early fusion.

Other features' combination methods involve a combination of late fusion and image re-ranking [12] (because the first step is the pre-filtering of the collection by text, the model is called semantic combination). Some researchers [9] experimented with tensoring of the representations and modeling the inherent dependencies between features' dimensions (although the incorporation of dependencies did not improve the retrieval effectiveness and the model was not scalable to large image collections due to its high computational cost).

The fusion approach that can be easily modified to incorporate the user feedback is based on the transmedia pseudo-relevance mechanism. This so-called inter-media feedback query expansion is based on textual query expansion in most of the papers ([13],[14]). Typically, textual annotations from the top visually-ranked images (or from a mixed run) are used to expand a textual query.

There is a proliferation of other models that utilize user feedback (mono-modal) in order to improve the retrieval. In this paper, however, we focus on the issue of combining the visual and textual features in the context of user feedback, therefore we are interested in hybrid approaches that combine the visual and textual features, and also hybrid approaches that combine them within the context of user feedback.

Our main contribution is the proposed model for combining visual and textual systems within the context of user feedback. The model was inspired by the expectation value of the measurement utilized in quantum mechanics and the tensor product of the density matrices of the systems (that results in a density matrix of the composite system). It was designed to capture both intra-relationships between features' dimensions (visual and textual correlation matrices) and inter-relationships between visual and textual representations (tensor product). The model provides a sound and natural framework to seamlessly integrate multiple feature spaces by considering them as a composite system, as well as a new way of measuring the relevance of an image with respect to a context by applying quantum-like measurement. It opens a door for a series of theoretically well-founded further exploration routes, e.g. by considering the interference among different features.

## 3 Combining Visual and Textual Features within the Context of User Feedback

Modern retrieval systems allow the users to interact with the system in order to narrow down the search. This interaction takes the form of implicit or explicit feedback. The representations of the images in the feedback set are often aggregated or concatenated (or co-occurrence matrices may be aggregated to represent i.e. probability distribution

matrix). The information extracted from the feedback set is utilized to expand the query or re-rank the top images returned in the first round of the retrieval.

Here, we are going to introduce our model for visual and textual systems' combination within the context of a user feedback. The proposed model was inspired by the measurement used in quantum mechanics, which is based on an expectation value, predicted mean value of the measurement

$$\langle A \rangle = tr\left(\rho A\right) \tag{1}$$

where $tr$ denotes the trace operator, $\rho$ represents a density matrix of the system and $A$ is an observable. We can also represent an observable $A$ as a density matrix (corresponding to the query or an image in the collection). For more information on the analogies between quantum mechanics and information retrieval the curious reader is referred to [17].

We are going to use the tensor operator $\otimes$ to combine the density matrices corresponding to visual and textual feature spaces. In quantum mechanics, the tensor product of density matrices of different systems represents a density matrix of the combined system (see [15]).

Thus, the proposed measurement is represented by

$$tr\left((M_1 \otimes M_2) \cdot \left(\left(a^T \cdot a\right) \otimes \left(b^T \cdot b\right)\right)\right) \tag{2}$$

where $M_1$, $M_2$ represent density matrices (co-occurrence matrices) of the query and images in the feedback set corresponding to visual and textual spaces respectively, $a$ and $b$ denote vectors representing visual and textual information for an image from the data collection, and $T$ is a transpose operation. We would perform this measurement on all the images in the collection, thus re-scoring the dataset based on the user feedback.

Assuming that the systems were prepared independently (otherwise we would have to try to model a concept analogous to entanglement [18]), we get

$$\begin{aligned} tr\left((M_1 \otimes M_2) \cdot \left(\left(a^T \cdot a\right) \otimes \left(b^T \cdot b\right)\right)\right) = \\ tr\left(\left(M_1 \cdot \left(a^T \cdot a\right)\right) \otimes \left(M_2 \cdot \left(b^T \cdot b\right)\right)\right) = \\ tr\left(M_1 \cdot \left(a^T \cdot a\right)\right) \cdot tr\left(M_2 \cdot \left(b^T \cdot b\right)\right) = \\ \left\langle M_1 | a^T \cdot a \right\rangle \cdot \left\langle M_2 | b^T \cdot b \right\rangle \end{aligned} \tag{3}$$

where $\langle \cdot | \cdot \rangle$ denotes an inner product operating on a vector space.

Let $q_v$, $q_t$ denote the visual and textual representations of the query, $c^i$, $d^i$ denote visual and textual representations of the images in the feedback set, $r_1$, $r_2$ denote the weighting factors (constant, importance of query and feedback density matrices respectively), and n denote the number of images in the feedback set. Then, we define $M_1$ and $M_2$ as weighted combinations of co-occurrence matrices (a subspace generated by the query vector and vectors from the feedback set)

$$\begin{aligned} M_1 = r_1 \cdot D_q^v + \frac{r_2}{n} \cdot D_f^v = \\ r_1 \cdot q_v^T \cdot q_v + \sum_i \left(\frac{r_2}{n} \cdot \left(c^i\right)^T \cdot c^i\right) \end{aligned} \tag{4}$$

and

$$M_2 = r_1 \cdot D_q^t + \frac{r_2}{n} \cdot D_f^t = \\ r_1 \cdot q_t^T \cdot q_t + \sum_i \left( \frac{r_2}{n} \cdot \left(d^i\right)^T \cdot d^i \right) \tag{5}$$

Co-occurrence matrices are quite often utilized in the Information Retrieval (IR) field. Because they are Hermitian and positive-definite, they can be thought of as density matrices (probability distribution). The common way of co-occurence matrix generation is to multiply the term-document matrix by its transpose (rows of the matrix represent the documents $d_1, \ldots d_m$), that is $D = M^T \cdot M$. Notice, that this is equivalent to $D = \sum_{i=1}^{n} d_i^T \cdot d_i$.

This observation, due to the properties of the inner product, will allow us to further simplify our model

$$\left\langle M_1 \otimes M_2 \right| \left(a^T \cdot a\right) \otimes \left(b^T \cdot b\right) \rangle = \left\langle M_1 | a^T \cdot a \right\rangle \cdot \left\langle M_2 | b^T \cdot b \right\rangle = \\ \left\langle r_1 \cdot q_v^T \cdot q_v + \sum_i \left( \frac{r_2}{n} \cdot \left(c^i\right)^T \cdot c^i \right) | a^T \cdot a \right\rangle \cdot \\ \left\langle r_1 \cdot q_t^T \cdot q_t + \sum_i \left( \frac{r_2}{n} \cdot \left(d^i\right)^T \cdot d^i \right) | b^T \cdot b \right\rangle = \\ \left( \left\langle r_1 \cdot q_v^T \cdot q_v | a^T \cdot a \right\rangle + \sum_i \frac{r_2}{n} \left\langle \left(c^i\right)^T \cdot c^i | a^T \cdot a \right\rangle \right) \cdot \\ \left( \left\langle r_1 \cdot q_t^T \cdot q_t | b^T \cdot b \right\rangle + \sum_i \frac{r_2}{n} \left\langle \left(d^i\right)^T \cdot d^i | b^T \cdot b \right\rangle \right) = \\ \left( r_1 \cdot \left\langle q_v | a \right\rangle^2 + \frac{r_2}{n} \cdot \sum_i \left\langle c^i | a \right\rangle^2 \right) \cdot \left( r_1 \cdot \left\langle q_t | b \right\rangle^2 + \frac{r_2}{n} \cdot \sum_i \left\langle d^i | b \right\rangle^2 \right) \tag{6}$$

Notice, that the model breaks down into the weighted combinations of individual measurements. The squares of the inner products come from the correlation matrices and can play an important role in the measurement. Later in the paper, we are going to justify this claim.

We can consider a variation of the aforementioned model, where just like in the original one $M_1 = r_1 \cdot D_q^v + \frac{r_2}{n} \cdot D_f^v$ and $M_2 = r_1 \cdot D_q^t + \frac{r_2}{n} \cdot D_f^t$. We can decompose (eigenvalue decomposition) the density matrices $M_1$, $M_2$ to estimate the bases[1] $(p_i^v, p_j^t)$ of the subspaces generated by the query and the images in the feedback set. Now, let us consider the measurement

$$\left\langle P_1 \otimes P_2 \right| \left(a^T a\right) \otimes \left(b^T b\right) \rangle \tag{7}$$

[1] It has been highlighted [19] that the orthogonal decomposition may not be the best option for visual spaces because the receptive fields that result from this process are not localized, and the vast majority do not at all resemble any known cortical receptive fields. Thus, in the case of visual spaces, we may want to utilize decomposition methods that produce non-orthogonal basis vectors.

where $P_1$, $P_2$ are the projectors onto visual and textual subspaces generated by query and the images in the feedback set ($\sum_i \left(p_i^v\right)^T p_i^v$, $\sum_j \left(p_j^t\right)^T p_j^t$), and $a$, $b$ are the visual and textual representations of an image from the data set. Because the tensor product of the projectors corresponding to visual and textual Hilbert spaces ($H_1$, $H_2$) is a projector onto the tensored Hilbert space ($H_1 \otimes H_2$), the measurement (7) can be interpreted as probability of relevance context, the probability that vector $a \otimes b$ was generated within the subspace (representing the relevance context) generated by $M_1 \otimes M_2$. Hence

$$\begin{aligned}
\left\langle P_1 \otimes P_2 \right| \left(a^T a\right) \otimes \left(b^T b\right) \rangle = \\
\left\langle P_1 | a^T a \right\rangle \cdot \left\langle P_2 | b^T b \right\rangle = \\
\left\langle \sum_i \left(p_i^v\right)^T p_i^v | a^T a \right\rangle \cdot \left\langle \sum_j \left(p_j^t\right)^T p_j^t | b^T b \right\rangle = \\
\sum_i \left\langle p_i^v | a \right\rangle^2 \cdot \sum_j \left\langle p_j^t | b \right\rangle^2 = \\
\sum_i Pr_i^v \cdot \sum_j Pr_j^t = \\
\left\| \left(\left\langle p_1^v | a \right\rangle, \ldots, \left\langle p_n^v | a \right\rangle\right) \otimes \left(\left\langle p_1^t | b \right\rangle, \ldots, \left\langle p_n^t | b \right\rangle\right) \right\|^2
\end{aligned} \tag{8}$$

where $Pr$ denotes the projection probability and $\|\cdot\|$ represents vector norm.

We can see, that this measurement is equivalent to the weighted combinations of all the probabilities of projections for all the images involved. In quantum mechanics, the square of the absolute value of the inner product between the initial state and the eigenstate is the probability of the system collapsing to this eigenstate. In our case, the square of the absolute value of the inner product can be interpreted as a particular contextual factor influencing the measurement.

In this paper, we are going to experimentally test the model based on the expectation value of the measurement and the tensor product of density matrices. The proposed model can incorporate both implicit (i.e. query history) and explicit (i.e. relevance data) forms of user feedback.

## 4 Experiments and Discussion

We evaluate the proposed model on ImageCLEFphoto 2007 data collection [20]. ImageCLEFphoto2007 consists of 20000 everyday real-world photographs. It is a standard collection used by Information Retrieval (IR) community for evaluation purposes. This allows comparison with published results. There are 60 query topics that do not belong to the collection.

Because of the abstract semantic content of many of the queries, ImageCLEFphoto 2007 data collection is considered to be very difficult for retrieval systems. For example, the topic “straight road in the USA” could be very difficult for visual features whereas “church with more than two towers” could render the textual features helpless. That is why the hybrid models should play an important role in modern retrieval systems.

### 4.1 Experimental Setup

We test our model (expectation value with a tensor product of density matrices) within a simulated user feedback framework. First, we perform the first round retrieval for a topic from the query set based on the visual features only (we retrieve 1000 images). We use the visual features only because in the real life scenario many images would not have textual descriptions. We also do not combine the features in the first round retrieval as this would represent a different task. In this work we want to focus on testing the features' combination models within the user feedback framework.

Next, we identify 1, 2 and 3 relevant images respectively from the highest ranked images based on the ground truth data (starting from the most similar). Thus obtained images simulate the user feedback and are utilized in the proposed model to re-score the data collection. For each query topic (60 in total) we calculate mean average precision (MAP) for the top 20 retrieved images, as most users would only look at this number of documents. We set the weights $r_1$, $r_2$ to 1 and $0.8$ respectively (standard weights' values for query and its context as in the classic Rocchio algorithm, for example).

The visual features used in the experiment are based on the Bag of Visual Words framework (see [21] for a detailed description). They are regarded as a mid-level representation.

The textual features were obtained by applying the standard Bag of Words technique, with Porter stemming, stop words removal, and term frequency - inverse document frequency weighting scheme.

### 4.2 Experimental Results and Discussion

As aforementioned, we modify existing models in order to incorporate the user feedback. We use several baselines for comparison purposes.

Thus, early fusion is represented by a modified Rocchio algorithm ($earlyFusion$). The only difference between this variation and the classic model is that we apply it to concatenated visual and textual vectors, as opposed to visual or textual representations only. Let $\oplus$ denote the concatenation operation (other notation as in the previous sections). Then, this model modify the query in a following way

$$newQuery = q_v \oplus q_t + \frac{0.8}{n} \sum_i (c_i \oplus d_i) \tag{9}$$

After the query modification the scores are recomputed.

Another baseline, which we will refer to as $lateFusion$ will be represented as a combination of all the scores

$$sim\,(q_v, a) + \frac{0.8}{n} \sum_i sim\,(c_i, a) + sim\,(q_t, b) + \frac{0.8}{n} \sum_i sim\,(d_i, b) \tag{10}$$

where $sim$ denotes the similarity between given vectors. In this work $sim$ is an inner product between two vectors.

Our third baseline $rerankText$ denotes the re-ranking of the results obtained from the first round retrieval based on the aggregated textual representations of the feedback

images. Similarly, $rerankVis$ represents re-ranking of the top retrieved images based on the aggregated visual representations of the images from the feedback set.

Next model $trMedia$ represents, as the label suggests, inter-media feedback query modification. Here, textual annotations from the feedback images (identified by visual features) are used to expand a textual query.

The system performance without simulated feedback will be denoted as $noFeedback$ and the proposed model for combination of visual and textual features within the context of simulated relevance feedback will be denoted as $prMeanMeasure$.

Table 1 presents the obtained results.

**Table 1.** Simulated Relevance Feedback, ImageCLEF2007photo results (MAP)

| | **1 Feedback Image** | **2 Feedback Images** | **3 Feedback Images** |
|---|---|---|---|
| $noFeedback$ | 0.013 | 0.013 | 0.013 |
| **prMeanMeasure** | **0.079** | **0.094** | **0.11** |
| $earlyFusion$ | 0.066 | 0.082 | 0.085 |
| $lateFusion$ | 0.066 | 0.082 | 0.085 |
| $rerankText$ | 0.055 | 0.069 | 0.075 |
| $rerankVis$ | 0.034 | 0.036 | 0.031 |
| $trMedia$ | 0.061 | 0.078 | 0.081 |

From the experimental results we can see, that the best performing model is based on the proposed predicted mean value of the measurement ($prMeanMeasure$) with the density matrix of the composite system (tensor product of the subspaces). The difference (in terms of means) between $prMeanMeasure$ and the rest of the baselines is statistically significant (paired t-test, $p < 0.05$). The inter-media feedback query expansion ($trMedia$) also performed well, albeit worse than early and late fusion ($earlyFusion$, $lateFusion$). In general, all the models' performance suggests that they are quite effective in utilizing users' feedback.

An interesting observation is that both early ($earlyFusion$, modified Rocchio) and late fusion strategies ($lateFusion$, combination of scores) show exactly the same performance. It is because

$$
\begin{aligned}
newQuery &= q_v \oplus q_t + \frac{0.8}{n}\sum_i (c_i \oplus d_i) \\
imagesInDataset &= a \oplus b \quad forAll \quad a, b \in Dataset \\
&\langle newQuery | imagesInDataset \rangle = \\
&\left\langle q_v \oplus q_t + \frac{0.8}{n}\sum_i (c_i \oplus d_i) \,|\, a \oplus b \right\rangle = \\
&\langle q_v \oplus q_t | a \oplus b \rangle + \frac{0.8}{n}\sum_i \langle c_i \oplus d_i | a \oplus b \rangle = \\
&\langle q_v | a \rangle + \frac{0.8}{n}\sum_i \langle c_i | a \rangle + \langle q_t | b \rangle + \frac{0.8}{n}\sum_i \langle d_i | b \rangle
\end{aligned}
\tag{11, 12}
$$

Thus, in our case the early and late fusion strategies (modified Rocchio algorithm operating on concatenated representations and weighted linear combination of scores) are interchangeable. We are going to address this interesting discovery in our future work.

We observe, that even one feedback image can help to narrow down the search, thus increasing the match between user's preferences (in this case, a human expert who assesed the relevance of images in ground truth data). Let us assume, that the visual query pictures a person wearing sunglasses. In the first round retrieval, the system may recognize (return more images of) a concept representing sunglasses without a person present on the picture. However, the human assesor might have deemed an image relevant only if both concepts were present in the image. A user feedback can then reinforce the subjective (perceived) relevance of the query to the retrieved images. In case of using the visual representations only in the user feedback ($rerankVis$), more images in the feedback set can sometimes confuse the visual features (especially if they significantly differ in terms of colour, texture, viewpoint or illumination). Thus, approaches like $rerankVis$ may strongly depend on the type of visual features used (while visual features A may be suitable for the particular feedback set C, visual features B may not work so well on C and vice versa).

In this work, the MAP is calculated for 20 top images only as this is a more realistic scenario (especially for user simulation/user feedback context). However, for 1000 top and 3 feedback images, the system performace is approximately $MAP \approx 0.206$. If we consider the ImageCLEF2007photo results of other systems (the best models utilize both visual and textual information) which can be found on the ImageCLEF website [23], the proposed model places itself among the best performing approaches. However, it must be noted that our model combines visual and textual features within the context of user feedback framework (different task).

We also need to take into consideration the disadvantages of automatic evaluation methods. The ultimate test for every retrieval system (especially for user simulation/user feedback context) should be the real user evaluation (although it is a time consuming task). The relevance of an image is a highly subjective concept and the automatic evaluation seems to fail to address this problem. Moreover, there is a glitch in the trec-eval evaluation software, that can bring the reported results into question. To be more specific, if some images obtain the same similarity score, they will be re-ordered by the software. The result is that two identical submissions may get different performance scores.

## 5 Conclusions

In this paper, we have presented the model for visual and textual features' combination within the context of user feedback. The approach is based on mathematical tools also used in quantum mechanics - the predicted mean value of the measurement and the tensor product of the density matrices, which represents a density matrix of the combined systems. It was designed to capture both intra-relationships between features' dimensions (visual and textual correlation matrices) and inter-relationships between visual and textual representations (tensor product). The model provides a sound and natural framework to seamlessly integrate multiple feature spaces by considering them as a

composite system, as well as a new way of measuring the relevance of an image with respect to a context by applying quantum-like measurement. It opens a door for a series of theoretically well-founded further exploration routes, e.g. by considering the interference among different features. It is easily scalable to large data collections as it is general and computationally cheap. The results of the experiment conducted on ImageCLEF data collection show the significant improvement over other baselines.

## 6 Future Work

The future work will involve testing different notions of correlation within the proposed framework (we can construct correlation matrices in such a way that they can be regarded as density matrices). In this paper, we incorporate document/image level correlations only. However, in case of textual representations, we can also experiment with Hyperspace Analogue to Language (HAL). In the aformentioned approach, the context is represented by a sliding window of a fixed size (while in document level correlation the context is represented by the whole document). We can also consider a visual counterpart to HAL, where a window of a fixed size (e.g. square, circular) is shifted from one instance of a visual word to another. Then, the number of instances of visual words that appear in the proximity of the visual word on which the window is centered can be calculated. In case of a dense sampling, the window would be shifted analogously to HAL in text IR. If the sparse sampling was utilized, however, the window would shift from one instance of a visual word to another.

## References

1. Zhao, R., Grosky, W.I.: Narrowing the semantic gap-improved text-based web document retrieval using visual features. IEEE Transactions on Multimedia 4, 189–200 (2002)
2. Ferecatu, M., Sahbi, H.: TELECOM ParisTech at Image Clef photo 2008: Bi-modal text and image retrieval with diversity enhancement. In: Working Notes of CLEF (2008)
3. Martinez-Fernandes, J.L., Serrano, A.G., Villena-Roman, J., Saenz, V.D.M., Tortosa, S.G., Castagnone, M., Alonso, J.: MIRACLE at ImageCLEF 2004. In: Working Notes of CLEF (2004)
4. Yanai, K.: Generic image classification using visual knowledge on the web. In: Proceedings of the 11th ACM International Conference on Multimedia, pp. 167–176 (2003)
5. Tjondronegoro, D., Zhang, J., Gu, J., Nguyen, A., Geva, S.: Integrating Text Retrieval and Image Retrieval in XML Document Searching. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 511–524. Springer, Heidelberg (2006)
6. Maillot, N., Chevallet, J.P., Valea, V., Lim, J.H.: IPAL Inter-media pseudo-relevance feedback approach to ImageCLEF 2006 photo retrieval. In: CLEF Working Notes (2006)
7. Rahman, M.M., Bhattacharya, P., Desai, B.C.: A unified image retrieval framework on local visual and semantic concept-based feature spaces. J. Visual Communication and Image Representation 20(7), 450–462 (2009)
8. Simpson, M., Rahaman, M.M.: Text and content-based approaches to image retrieval for the ImageCLEF 2009 medical retrieval track. In: Working Notes for the CLEF 2009 Workshop (2009)

9. Wang, J., Song, D., Kaliciak, L.: Tensor product of correlated text and visual features: a quantum theory inspired image retrieval framework. In: AAAI-Fall 2010 Symposium on Quantum Information for Cognitive, Social, and Semantic Processes, pp. 109–116 (2010)
10. Mensink, T., Csurka, G., Perronnin, F.: LEAR and XRCE's participation to visual concept detection task - ImageCLEF 2010. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, pp. 77–80 (2006)
11. Mensink, T., Verbeek, J., Csurkay, G.: Weighted transmedia relevance feedback for image retrieval and auto-annotation. Technical Report Number 0415 (2011)
12. Clinchant, S., Ah-Pine, J., Csurka, G.: Semantic combination of textual and visual information in multimedia retrieval. In: ACM International Conference on Multimedia Retrieval, ICMR (2011)
13. Depeursinge, A., Muller, H.: Fusion techniques for combining textual and visual information retrieval. In: ImageCLEF. The Springer International Series on Information Retrieval, vol. 32, pp. 95–114 (2010)
14. Chang, Y.-C., Chen, H.-H.: Increasing Precision and Diversity in Photo Retrieval by Result Fusion. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 612–619. Springer, Heidelberg (2009)
15. Combining systems: the tensor product and partial trace, `http://www.quantum.umb.edu/Jacobs/QMT/QMT-AppendixA.pdf`
16. Li, Y., Cunningham, H.: Geometric and quantum methods for information retrieval. SIGIR Forum 42(2), 22–32 (2008)
17. van Rijsbergen, C.J.: The geometry of information retrieval. Cambridge University Press (2004)
18. Bruza, P.D., Kitto, K., Nelson, D., McEvoy, C.L.: Entangling words and meaning. In: Proceedings of the 2nd Quantum Interaction Symposium, pp. 118–124 (2008)
19. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609 (1996)
20. Grubinger, M., Clough, P., Hanbury, A., Muller, H.: Overview of the ImageCLEF 2007 photographic retrieval task. In: Working Notes of the 2007 CLEF Workshop (2007)
21. Kaliciak, L., Song, D., Wiratunga, N., Pan, J.: Novel local features with hybrid sampling technique for image retrieval. In: Proceedings of Conference on Information and Knowledge Management (CIKM), pp. 1557–1560 (2010)
22. Nowak, E., Jurie, F., Triggs, B.: Sampling Strategies for Bag-of-Features Image Classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 490–503. Springer, Heidelberg (2006)
23. ImageCLEF website, `http://www.imageclef.org`
24. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEFphoto 2007 Photographic Retrieval Task. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 433–444. Springer, Heidelberg (2008)
25. Chen, Z., Liu, W., Zhang, F., Li, M.J., Zhang, H.J.: Web mining for web image retrieval. Journal of the American Society for Information Science and Technology 52(10), 831–839 (2001)