Content-Based Image Retrieval in Augmented Reality

Leszek Kaliciak(✉), Hans Myrhaug, and Ayse Goker

Ambiesense Ltd., Aberdeen, UK
{leszek, hans, ayse}@ambiesense.com

Abstract. In this paper, we present a content-based image retrieval framework which augments the user's reality and supports the decision making process as well as awareness and understanding of the local marine environment. It comprises a real-time intelligent user interface combined with the 360° real-time environment display in the virtual reality headset. The image retrieval utilizes a unified hybrid adaptive image retrieval model. The presented system provides the user with a unique solution combining the virtual reality real-time headset, 360° view, and augmented reality to remotely monitor the surface and underwater marine environment. The objective of the proposed framework is to enhance the user interaction with the remote sensing and control applications. To our knowledge, it is the first system that combines real-time VR, 360° camera, and hybrid models in the context of image retrieval and augmented reality.

Keywords: Virtual reality · Augmented reality · 360° cameras · Content-based image retrieval · Hybrid models

1 Introduction and Related Work

Augmented reality (AR) is a live view of a physical, real-world environment augmented by computer-generated information. It has applications to navigation, commerce, captioning, among others [9]. Augmented reality is especially well suited for the environment surveillance purposes, for example surface and underwater ocean monitoring. Immersive virtual reality (VR) is often not considered augmented reality because the user is not exposed to a real but a virtual world. We, however, use the virtual reality headset to directly stream the real world information in real time and augment it with computer generated content. The presented use-case is a part of the bigger project which aims at developing a design environment for Cyber Physical Systems. The new type of marine robot with surface and underwater surveillance capabilities is presented in Fig. 1.

In terms of related work, a 360° camera with the virtual reality headset was also used in [10] to provide the users with novel and intuitive means to view the imagery of the Antarctic Ocean from a vessel. By applying the alignment and

J.F. De Paz et al. (eds.), *Ambient Intelligence—Software and Applications—8th International Symposium on Ambient Intelligence (ISAmI 2017)*,
Advances in Intelligent Systems and Computing 615, DOI 10.1007/978-3-319-61118-1_13

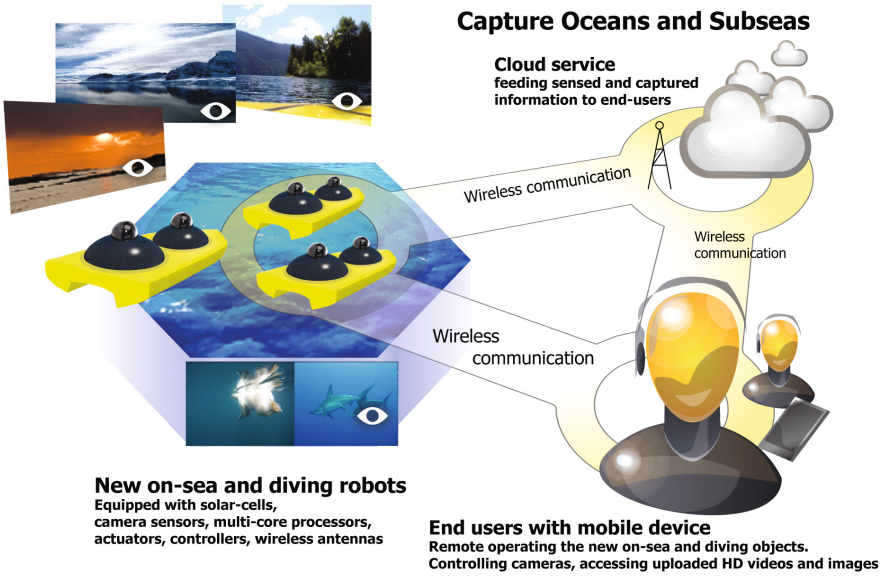

Fig. 1. The marine robots - new cyber-physical systems with immersive remote sensing capabilities

re-projection technique, the 3D image of the vessel's environment is produced in Unreal Engine. This allows for a seamless movement in 3 dimensions.

Another related system taking advantage of the 360° real-time camera and VR is the one introduced in [2]. The presented idea is a vision of the augmented reality real-time 3D environment for robotic operations in space. The astronauts, with the help of VR headset, would be able to quickly and easily move through a virtual reality enhanced 3D model of the International Space Station validated and augmented with real-time camera video.

Our marine robot augments the real-time surface and underwater data stream with the information about similar (previously encountered) information objects. The retrieval of similar objects is based on the fusion of different types of features in order to reduce the semantic gap, the difference between machine representation and human perception of images.

The main challenges in data fusion are currently related to incorporation of correlation and adaptivity into hybrid models [11]. Different feature spaces are correlated because they often represent the same information object. Moreover, the importance of different feature spaces varies, it is dynamic in nature. Standard widely used data fusion strategies (e.g. linear combination of scores) do not take this correlation into account and use fixed weights associated with the importance of feature spaces [1,7].

The presented framework integrates our prototype unified tensor-based system consisting of the user interface, the hybrid model for the combination of visual and textual features, and the hybrid adaptive model for the combination of

features in the context of relevance feedback. We address the main challenges in data fusion by utilizing the inherent visual-textual (inter) and visual-visual, and textual-textual (intra) relationships between visual and textual feature spaces. Moreover, the hybrid relevance feedback model adapts its weights associated with the importance of visual and textual features based on the interactions with the user - user relevance feedback.

2 Augmented Reality User Interface

One of the forms of interaction of the user with the robot is via the virtual reality (VR) headset Fig. 2. The VR headset displays the surface and underwater 3D environment captured by the 360° spherical cameras mounted on the marine robots. The experience becomes even more immersive with the head-tracking like feature. All of this offer the user a natural, real-time marine monitoring capability. Moreover, the real-time visual image of the marine environment is augmented by additional digital information. When using the select button on the VR control pad, the on-the-fly image capture and retrieval presents the user with the visual and textual information related to similar information objects. A pop-up window is used to display the retrieval results which can be freely browsed. The user can further narrow down the presented top retrieval results by highlighting the relevant images with one of the control pad buttons and pressing the select button. When using the select button on one of the top result images, the associated textual information will be displayed.

The augmented reality user interface is shown in Fig. 3.

Fig. 2. Marine robot engineer accessing the remotely sensed imagery in VR glasses

Fig. 3. The augmented reality user interface

3 Implemented Models

In this section we present the advanced hybrid image retrieval models used in conjunction with the augmented reality VR headset.

Our models combine different types of visual information representing a specific information object (image), and also textual and visual information for the search refinement process providing the textual information is available in the database of similar images.

3.1 Image Representation

The visual features implemented in our prototype model comprise global and local methods: edge histogram, homogeneous texture, bag of visual words features, colour histogram, and co-occurrence matrix. The textual features used in the search refinement process are based on the standard vector space model with the tf-idf weighting scheme.

In addition, the extracted visual representations of the information objects incorporate the absolute spatial information in order to focus on the middle part of the 360° images. Because the user has the freedom to look around the environment in the VR headset, he/she will focus his/her gaze on the most interesting areas of the 360° image.

In order to implement the aforementioned absolute spatial information, we divide an image into two regions. The central circular region represents the user's focus, and the remaining region represents his/her peripheral vision. We use both regions in the retrieval process by extracting the same visual features from each region and concatenating the representations.

3.2 Hybrid Adaptive Image Retrieval

The data used in Multimedia Retrieval is often multimodal and heterogeneous. Tensors, as generalizations of vectors and matrices, provide a natural and scalable framework for handling data with inherent structures and complex dependencies. There has been a recent renaissance of tensor-based methods in machine learning. They range from scalable algorithms for tensor operations, novel models through tensor representations, to industry applications such as Google TensorFlow, Torch and Tensor Processing Unit (TPU).

We have developed and implemented two tensor-based hybrid models in our prototype system: a hybrid model for the combination of visual and textual features [4], and a hybrid adaptive relevance feedback model for the combination of features in the context of user relevance feedback [5].

In the first tensor-based model we utilize a specific combination of the Euclidean distance to measure the similarity between visual representations, and the cosine similarity to measure the similarity between textual representations. The Euclidean distance, in the case of our mid-level visual features, performs better than cosine similarity. It is due to the fact that normalization of our local features hampers the retrieval performance. On the other hand, cosine similarity in textual space performs better than other similarity measurements [6]. This specific combination of scores has an interesting interpretation in the form of the Euclidean distance calculated on tensor-ed representations [4].

Thus, we combine the distances as

$$\sqrt{s_e^2\left(d_1^v, d_2^v\right) s_c\left(d_1^t, d_2^t\right) - 2 s_c\left(d_1^t, d_2^t\right) + 2} \tag{1}$$

where s_e denotes the Euclidean distance, s_c represents the cosine similarity measure, d_1^v and d_1^t denote visual and textual representations of the query respectively, d_2^v and d_2^t denote visual and textual representations of an image in the data collection respectively, and $\otimes$ is the tensor operator.

It can be shown that the aforementioned combination of measurements is equivalent to

$$\sqrt{s_e^2\left(d_1^v, d_2^v\right) s_c\left(d_1^t, d_2^t\right) - 2 s_c\left(d_1^t, d_2^t\right) + 2} = s_e\left(d_1^v \otimes d_1^t, d_2^v \otimes d_2^t\right) \tag{2}$$

Thus, the implemented model is equivalent to computing the Euclidean distance on tensor-ed representations.

The second tensor-based implemented hybrid model is the hybrid relevance feedback for image refinement. It utilizes the correlation and complementarity between different feature spaces and has been extensively evaluated against other state-of-the-art models [5]. Moreover, because query can be correlated with its context to a different extent [3,8], the implemented model adapts its weights based on the user relevance feedback.[1]

[1] In this paper, the textual and visual terms refer to image tags and instances of visual words, respectively.

The model is defined on a Hilbert space (a complex space with an inner product) which can be thought of as a natural extension of the standard vector space model, with its useful notions of subspaces and projections. It was inspired by the mathematical tools utilized in Quantum Mechanics (QM) and is based on the expectation value, the predicted mean value of the measurement. The model also uses the notions of co-occurrence and the tensor operation. Co-occurrence matrices can be treated as density matrices (probability distribution) because they are Hermitian and positive-definite, and the tensor operator can be utilized to combine the density matrices corresponding to visual and textual feature spaces. The tensor product of density matrices of different systems represents a density matrix of the combined system.

Thus, the intra-feature correlations are captured by density matrices corresponding to individual feature spaces, and inter-correlations are modeled in the form of the tensor product - resulting in a density matrix of the composite system. The projection of the query onto the subspace of the composite system can then be considered as our similarity measurement.

Let tr denotes the matrix trace operator, $\langle\cdot|\cdot\rangle$ represents an inner product, M_1, M_2 are co-occurrence matrices corresponding to different feature spaces (a subspace generated by the query vector and vectors from the feedback set), $\otimes$ denotes the tensor operator, a, b are different representations of an image in the collection corresponding to M_1 and M_2, 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, D_q^v, D_f^v denote the density (co-occurrence) matrices of the visual query and its visual context (feedback images), D_q^t, D_f^t denote the density matrices of the textual query and its textual context, and n denotes the number of images in the feedback set.

$$\begin{aligned} tr\left((M_1 \otimes M_2) \cdot \left(\left(a^T a\right) \otimes \left(b^T b\right)\right)\right) = \\ \left(str_v \left\langle q_v|a\right\rangle^2 + (1 - str_v)\frac{1}{n}\sum_i \left\langle c^i|a\right\rangle^2\right) \cdot \\ \left(str_t \left\langle q_t|b\right\rangle^2 + (1 - str_t)\frac{1}{n}\sum_i \left\langle d^i|b\right\rangle^2\right) \end{aligned} \tag{3}$$

where

$$str_v = \frac{\left\langle D_q^v \middle| D_f^v \right\rangle}{\left\|D_q^v\right\| \left\|D_f^v\right\|} = \frac{\sum_i \left\langle q_v|c^i\right\rangle^2}{\langle q_v|q_v\rangle \sqrt{n\sum_i \left\langle c^i|c^i\right\rangle^2}} \tag{4}$$

and

$$str_t = \frac{\left\langle D_q^t \middle| D_f^t \right\rangle}{\left\|D_q^t\right\| \left\|D_f^t\right\|} = \frac{\sum_i \left\langle q_t|d^i\right\rangle^2}{\langle q_t|q_t\rangle \sqrt{n\sum_i \left\langle d^i|d^i\right\rangle^2}} \tag{5}$$

Let us assume that the relevance feedback is provided after the first round retrieval to refine the query. The adaptive weighting can be interpreted in a following way:

1. small $\langle D_q | D_f \rangle$; weak relationship between query and its context, context becomes important. We adjust the probability of the original query terms; the adjustment will significantly modify the original query.
2. big $\langle D_q | D_f \rangle$; strong relationship (similarity) between query and its context, context will not help much. The original query terms will tend to dominate the whole term distribution in the modified model. The adjustment will not significantly modify the original query.

The adaptive model can be naturally expanded to accommodate other features, e.g. various visual features

$$tr\left(\left(\otimes_n M_n\right) \cdot \left(\otimes_n \left(a_n^T \cdot a_n\right)\right)\right) = \prod_n \left\langle M_n \middle| a_n^T \cdot a_n \right\rangle \tag{6}$$

Thus, for 3 features (e.g. two visual and a textual feature) the adaptive model becomes

$$\begin{aligned} tr\left(\left(M_1 \otimes M_2 \otimes M_3\right)\left(\left(a_1^T a_1\right) \otimes \left(a_2^T a_2\right) \otimes \left(b^T b\right)\right)\right) = \\ \left(str_{v1} \left\langle q_{v1} | a_1 \right\rangle^2 + \left(1 - str_{v1}\right) \frac{1}{n} \sum_i \left\langle c_1^i \middle| a_1 \right\rangle^2\right) \cdot \\ \left(str_{v2} \left\langle q_{v2} | a_2 \right\rangle^2 + \left(1 - str_{v2}\right) \frac{1}{n} \sum_i \left\langle c_2^i \middle| a_2 \right\rangle^2\right) \cdot \\ \left(str_t \left\langle q_t | b \right\rangle^2 + \left(1 - str_t\right) \frac{1}{n} \sum_i \left\langle d^i \middle| b \right\rangle^2\right) \end{aligned} \tag{7}$$

Here, for example, M_1, a_1 and M_2, a_2 may correspond to different visual features (density matrices and vector representations of images from the database), and M_3, b corresponds to the textual feature.

4 Conclusions and Future Work

In this paper we have presented the integrated ocean monitoring system for surface and underwater ocean surveillance. The main objective of the proposed framework was to enhance user interaction with remote sensing and control applications. The system offers the user an immersive monitoring experience by combining the augmented reality and the VR real-time real-life 360° view.

The augmented reality incorporates the hybrid image search model to retrieve similar images from the database of previously encountered marine environments to provide the user with supporting information.

The discussed unified tensor-based hybrid retrieval system comprises combinations of various visual features, combination of visual and textual feature spaces, combination of visual and textual feature spaces in the context of search refinement, and the user interface. The implemented models address the key challenges in data fusion, specifically correlation and adaptivity.

To our knowledge, the proposed framework is the very first to combine real-time VR, 360° camera, and hybrid models in the context of image retrieval and augmented reality.

Although our current efforts are focused on the marine robot, another interesting application will be the adaptation of the presented framework to the Mars rover use case which is also a part of the large R&D project. In that scenario, the 360° spherical cameras can be used to take pictures of Martian environment which can be later viewed in the VR headset with the help of additional search functionality.

Acknowledgements. This work has been partially funded by the CERBERO project no. 732105—a HORIZON 2020 EU project.

References

1. Bhowmik, N., Gonzalez, V.R., Gouet-Brunet, V., Pedrini, H., Bloch, G.: Efficient fusion of multidimensional descriptors for image retrieval. In: IEEE International Conference on Image Processing, pp. 5766–5770 (2014)
2. Goecks, V.G., Chamitoff, G.E., Borissov, S., Probe, A., McHenry, N.G., Cluck, N., Paddock, E., Schweers, J.P., Bell, B.N., Hoblit, J.: Virtual reality for enhanced 3D astronaut situational awareness during robotic operations in space. In: AIAA Information Systems-AIAA Infotech@ Aerospace, p. 0883 (2017)
3. Goker, A., Myrhaug, H., Bierig, R.: Context and information retrieval. In: Information Retrieval: Searching in the 21st century. Wiley (2009)
4. Kaliciak, L., Myrhaug, H., Goker, A., Song, D.: On the duality of fusion strategies and query modification as a combination of scores. In: The 17th International Conference on Information Fusion (Fusion 2014), Salamanca, Spain (2014)
5. Kaliciak, L., Myrhaug, H., Goker, A., Song, D.: Adaptive relevance feedback for fusion of text and visual features. In: The 18th International Conference on Information Fusion (Fusion 2015), pp. 1322–1329, Washington DC, USA (2015)
6. Liu, H., Song, D., Rueger, S., Hu, R., Uren, V.: Comparing dissimilarity measures for content-based image retrieval. In: The 4th Asia Information Retrieval Symposium, pp. 44–50 (2008)
7. Risojevic, V., Babic, Z.: Fusion of global and local descriptors for remote sensing image classification. IEEE Geosci. Remote Sens. Lett. **10**(4), 836–840 (2013)
8. Teevan, J., Dumais, S., Horvitz, E.: Personalizing search via automated analysis of interests and activities. In: 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 449–456 (2005)
9. Singh, M., Singh, M.P.: Augmented reality interfaces. IEEE Internet Comput. **17**(6), 66–70 (2013)

10. Sorensen, S., Kolagunda, A., Mahoney, A.R., Zitterbart, D.P., Kambhamettu, C.A.: Virtual reality framework for multimodal imagery for vessels in polar regions. In: MultiMedia Modeling: 23rd International Conference, MMM 2017, pp. 63–75 (2017)
11. http://fusion2015.org/plenary-speakers/