Automatic Selection of Web Contents Towards Automatic Authoring of a Video Biography

Ichiro Ide*, Yasutomo Kawanishi*, Kyoka Kunishiro†¶, Frank Nack‡, Daisuke Deguchi§*, Hiroshi Murase*

*Graduate School of Informatics, Nagoya University, Nagoya, Japan, 464–8601
Email: {ide, kawanishi, murase}@i.nagoya-u.ac.jp
†Graduate School of Information Science, Nagoya University, Nagoya, Japan, 464–8601
Email: kunishirok@murase.m.is.nagoya-u.ac.jp
‡Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands, 1098 XH
Email: nack@uva.nl
§Information Strategy Office, Nagoya University Nagoya, Japan, 464–8601
Email: ddeguchi@nagoya-u.jp
¶Currently at TOYOTA MOTOR CORPORATION.

***Abstract*—In this paper, we propose a method for image selection using Web image search for automatic video biography authoring. In the proposed method, images are selected from the image search results considering their visual contents for inclusion in the video biography. Through evaluation, we confirmed the effectiveness of the proposed image selection method compared to a baseline method which simply selects the top 1 search result.**

I. Introduction

Among various genres of news, the most popular ones are "sports" and "celebrity". Since these news involve famous people, their characters are often introduced in the form of video biographies which summarize their lives. Since a video biography visually introduces the character of a person-in-focus, it is usually more intuitive than text-based information. Thus, even if we are not familiar with the person, it would be easier to get acquainted with his/her character.

In reality, such a video biography is broadcasted after sudden incidents; mostly the death of the person introduced in the form of an obituary, which requires a speedy and timely authoring process. However, since manual editing of video is time consuming, the contents of the video biography tend to be composed of limited materials of well-known topics in recent years that are easily retrieved from the archive.

Meanwhile, Web portals such as Wikipedia[1] and image search engines have become popular and useful. Since they are sources of rich materials on famous people, we have been considering the usage of such materials for editing a video biography of a famous person.

We are currently implementing a framework that automatically authors a video biography using Web contents by the process-flow shown in Figure 1. First, given the name of a person-in-focus, the framework accumulates from an Wikipedia article, key-phrases that represent specific events concerning the person.

[1]https://en.wikipedia.org/wiki/

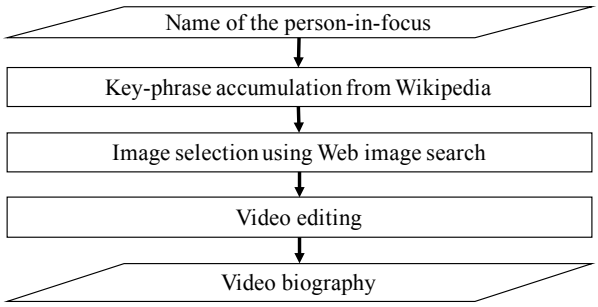


Fig. 1. Process-flow of the automatic video biography authoring framework.

Next, given a key-phrase and the name of the person, image search is performed. In general, it is considered that the top 1 search result is the most relevant to the search query [1]. However, in reality, the visual contents of the top 1 search result does not exactly match the query, even though it may be relevant in one way or another. This does not always satisfy our purpose, since it is important to select the most typical image related to the key-phrase. Therefore, in this paper, we propose an image selection method based on matching with general images of the given key-phrase to select the most visually suitable image among the top-ranked search results.

Finally, the selected images are concatenated into a video biography in the form of a slide-show.

In this paper, we focus on the method for image selection using Web image search, where images are selected from the image search results considering their visual contents for inclusion in the video biography.

Since most existing automatic video authoring methods [2], [3] mainly use given contents to edit a video, their users need to accumulate the contents beforehand, whereas the proposed method includes the contents accumulation process. With this method, we expect to realize efficient and effective authoring

978-1-5386-2937-6/17 $31.00 © 2017 IEEE
DOI 10.1109/ISM.2017.54

of, where a user only needs to input the name of the person-in-focus to obtain a video biography.

As few related works regarding contents accumulation, Liu et al. have proposed an automatic image suggestion method for presentation purposes [4]. However, since their work aims at supporting manual authoring of presentation materials, it considers "semantic diversity" and "visual diversity" among multiple suggestions proposed to users as important criteria. This is different from our criterion where selecting the most suitable (or in other words, typical) image that matches a given query is the important factor.

The rest of this paper is organized as follows: In Section II, we describe the proposed key-phrase accumulation method. In Section III, we describe the proposed image selection method. In Section IV, we report the result of an experiment. Finally, we conclude the paper in Section V.

II. Key-phrase Accumulation

Given the name of a person-in-focus, we accumulate key-phrases from an Wikipedia article on the person so that they could be used as text queries for image search.

By manually analyzing obituaries actually broadcasted on television (BBC (UK), NHK (Japan), NPO (The Netherlands), RTL(Germany), and WDR (Germany)), we concluded that in order to summarize the life of a famous person, a video biography should be composed of three topics; *childhood*, *profession*, and *personal life*.

For the key-phrases concerning *childhood*, we accumulate typical information from DBpedia [2] [5] which is a structured database of information obtained from Wikipedia.

Since the type of information concerning *profession* and *personal life* are different per person and thus not structured and described uniformly in DBpedia, we accumulate keyphrases concerning these topics directly from an Wikipedia article on the person. In Wikipedia, editors can emphasize phrases by using the Wiki syntax. Since the phrases tend to be characteristic information, referring to Wiki syntax, we accumulate them as key-phrases concerning the *profession* and the *personal life* of the person. Since the first block of an Wikipedia article is likely to be a summary that contains the most representative information in the entire article, we primarily accumulate key-phrases from this block. To classify a key-phrase into either of the two topics after accumulation, we apply word template matching to the titles of all the subsections in the article.

We additionally accumulate key-phrases concerning *personal life* from the remaining part of the Wikipedia article. We defined five sub-topics under the *personal life* topic: *hobby*, *side-business*, *social activism*, *spouse*, and *work*, and prepared word templates to detect additional key-phrases for each of them. These sub-topics were defined by analyzing video biographies actually broadcasted on television.

An extracted key-phrase which is a section title of the Wikipedia article and identical to a templates is not suitable for a key-phrase since it will be too general. In this case, we decided to substitute it to make it more concrete.

A substitute key-phrase is accumulated from the section corresponding to the title identical to the template. As same as the method introduced in the previous paragraph, emphasized phrases which match to word templates corresponding to the sub-topics are extracted as the substitute key-phrases. Note that, in order to increase the variety of templates for this purpose, we expanded them by accumulating terms cooccurring to those in the templates using Word2Vec [6].

[2]http://dbpedia.org/sparql/

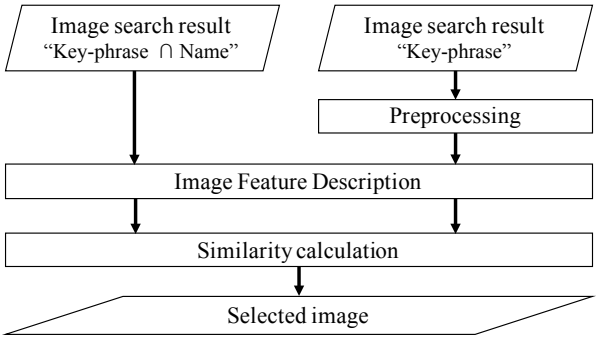


Fig. 2. Process-flow of the image selection using a key-phrase and the name of the person-in-focus.

III. Image Selection

The method searches images using a combination of key-phrase q and the name p of the person-in-focus as a query, and sort them by search ranking. We named the top M images as "Personalized Key-Phrase Images" I^P. Formally,

$$I^P = S(q \cap p) = \{i_1^P, i_2^P, \ldots, i_M^P\}, \tag{1}$$

where the function $S(q \cap p)$ returns images i_m^P $(m = 1, 2, \ldots, M)$ retrieved by a joint query q and p.

We want to select an image whose contents match the key-phrase best. For that, the method searches images using only key-phrase q as a query, and sort them by the search ranking. We named the top N images as "General Key-phrase Images" I^G. Formally,

$$I^G = S(q) = \{i_1^G, i_2^G, \ldots, i_N^G\}, \tag{2}$$

where the function $S(\cdot)$ returns images i_n^G $(n = 1, 2, \ldots, N)$ retrieved by an image search engine.

By comparing each image $i_m^P \in I^P$ with the General Key-phrase Images I^G, the most suitable image $\tilde{i}^P$ for key-phrase q of the person-in-focus p is selected as follows:

$$\tilde{i}^P = \arg \max_{i_n^P \in I^P} f(i_n^P, I^G), \tag{3}$$

where the similarity function $f(\cdot, \cdot)$ is switched according to the topic of the key-phrase. The process-flow is shown in Figure 2.

In the case of the key-phrases concerning *profession*, since General Key-phrase Images tend to contain a specific object

such as a "gold medal", the similarity function $f(\cdot,\cdot)$ is measured using local features. On the other hand, in the case of key-phrases concerning *personal life*, since the contents of General Key-phrase Images tend to be relatively ambiguous, the similarity $f(\cdot,\cdot)$ is measured using Visual Concepts [7], [8], which can represent the target images in a more abstract fashion. Note that in the case of the key-phrases concerning *childhood*, we did not apply the image selection process since currently we only searched for an image of the birthplace (usually a place name) of the person-in-focus; The top 1 search result was selected in this case.

A. Image Selection Using Local Features

If the topic of a given key-phrase q is *profession*, the most suitable image $\tilde{i}^P$ is selected by comparing the similarity of local features.

Although I^G is a set of top-ranked images of the image search result, they often contain images irrelevant to the query. Therefore, the method filters out the irrelevant images by applying mean-shift clustering [9].

Before the clustering, Principal Component Analysis (PCA) [10] is applied to 2,500 dimensional vectors of scaled-down 50 pixels square images in I^G. These images I^G are compressed to a dimension so that its cumulative contribution ratio should exceed a threshold, and finally, we obtain filtered General Key-phrase Images $\tilde{I}^G$. We use Oriented Fast and Rotated Brief (ORB) [11] for calculating the similarity between image i_n^P and the filtered images $\tilde{I}^G$.

As a similarity metric, we use the number of matched keypoints between the image i_n^P and each of the filtered image $\tilde{i}^G \in \tilde{I}^G$.

B. Image Selection Using Visual Concepts

If the topic of a given key-phrase q is *personal life*, the most suitable image $\tilde{i}^P$ is selected by comparing the similarity of the distribution of Visual Concepts detected in the image. Note that before the image selection, since color information is one of the most important cues for detecting Visual Concepts accurately, Automatic Image Colorization (AIC) [12] is applied if images in the search results are in gray scale.

Using Visual Concept detectors, likelihood for each Visual Concept in an image could be obtained. Here, we consider that the top U Visual Concepts according to their likelihood should represent an image. Since the appearance frequency of each Visual Concept is different among concepts, we weight each likelihood in a tf-idf [13] manner.

Since the likelihood histogram becomes very sparse only using the top U Visual Concepts, we expand the vocabulary with co-occurring terms obtained using Word2Vec. Using Word2Vec, cooccurring terms of the names of the top U Visual Concepts can be obtained. For each cooccurring term, its likelihood is calculated by multiplying the likelihood of the corresponding Visual Concept and the cooccurrence probability of the respective cooccurring term. Each image $i_m^P \in I^P$ $(m = 1, 2, \ldots, M)$ is described by using an expanded histogram of Visual Concepts.On the other hand, General Key-phrase Images I^G are described by an expanded vector of the sum of all histograms of Visual Concepts extracted from all images $i_n^G \in I^G$ $(n = 1, 2, \ldots, N)$. Let $\mathbf{h}_{m'}$ and $\mathbf{h}$ denote the L_1-normalized expanded histograms of Visual Concepts calculated from an image $i_{m'}^P$ and that of I^G, respectively. The similarity function $f(\cdot,\cdot)$ is then defined as follows:

$$f(i_{m'}^P, I^G) = 1 - \frac{(\mathbf{h} - \mathbf{h}_{m'})^2}{\max_m (\mathbf{h} - \mathbf{h}_m)^2}, \tag{4}$$

where $\mathbf{h}_m$ is an expanded histogram of Visual Concepts extracted from image i_m^P.

IV. Evaluation

A. Evaluation of the Image Selection Method

We evaluated the image selection process and confirmed its effectiveness. For the evaluation, we constructed a dataset that consists of 32 key-phrases accumulated from the 29 persons whose professions were *expert*, *public figure*, *artist*, or *athlete*.

For each key-phrase, four images were retrieved by Microsoft's Bing image search[3]. These sets of four images were used as the dataset.

The ground-truth was determined by subjective evaluation, in which twelve male Computer Science major students in their twenties participated. For each set of four retrieved images, each subject selected one that he considered the most appropriate to represent the query. As a result, the ratio of subjects who considered the image as most appropriate (hereafter, appropriateness ratio) was given to each image as the ground-truth.

We selected one image out of each set of four retrieved images in the dataset by the proposed and the comparative methods. The results were evaluated as the average appropriateness ratio across all key-phrases. Here, if we assumed that an image with the highest appropriateness ratio should be selected, the accuracy of the proposed method was 65.9%.

For comparison, we prepared methods that used SIFT [14] or AKAZE [15] instead of ORB as local feature. We also prepared methods that do not employ the feature switching according to the key-phrase topic; One method measured the similarity using only local feature (ORB) for key-phrases concerning both *profession* and *personal life*, while another one measured using only Visual Concepts in both cases. As a baseline, we prepared a simple method that automatically selects the top 1 image search result. Note that for the Visual Concept detection, we used the GoogLeNet detector that could detect 1,000 Visual Concepts [16].

The result of the evaluation is shown in Table I. The proposed method achieved the highest appropriateness ratio of 38.8%. From this result, we configured that the feature switching according to the key-phrase topic was effective and also ORB was the best local feature to be used.

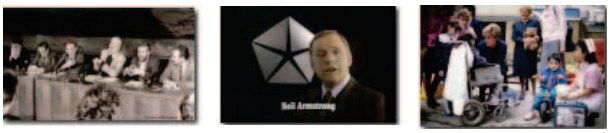
Examples of successfully selected images by the proposed method are shown in Figure 3. In the case of Figure 3(a), since

[3]https://www.bing.com/

TABLE I
EVALUATION OF THE IMAGE SELECTION METHOD

Method	Average appropriateness ratio
Proposed (with ORB and per-topic features)	**38.8%**
Proposed with AKAZE instead of ORB	27.8%
Proposed with SIFT instead of ORB	25.4%
Proposed with local features (ORB) for both topics	36.8%
Proposed with Visual Concepts for both topics	28.7%
Baseline by selecting the top 1 search result	26.6%

(a) Topic *profession*



(b) Topic *personal life*

Fig. 3. Examples of images successfully selected by the proposed method.

the topic of the given key-phrase was *profession*, local feature (ORB) was used for the similarity measurement. Meanwhile, in the case of Figure 3(b), since the topic of the given key-phrase was *personal life*, Visual Concepts were used for the similarity measurement.

V. CONCLUSION

In this paper, we proposed an automatic selection method of Web contents for automatic video biography authoring, which only requires the name of a person-in-focus as input. Given a key-phrase and the name of the person, the image selection process selects an image from Web image search results, whose visual contents match the key-phrase best.

Future work includes the following items in addition to the implementation and the evaluation of the key-phrase accumulation and video authoring processes.

- Increasing the number and variety of Visual Concepts that can be detected.
- Expanding the proposed image selection process to videos. This could be done with the current method, but we need to design an efficient method to deal with a large number of consecutive images that compose a video.
- Refining the image selection process. Although the main criteria are different between our work and Liu et al.'s work [4] introduced in Section I, their criterion on "visual quality" could also be considered in our method.

VI. ACKNOWLEDGMENTS

We would like to thank the subjects who participated in the experiments. Parts of this research were supported by a joint research project with NII, Japan, by the MEXT Grant-in-Aid for Scientific Research, and the JSPS Invitation Fellowship for Research in Japan (Short-term).

REFERENCES

[1] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proceedings of the Sixth ACM International Conference on Image and Video Retrieval*. Amsterdam, The Netherlands: ACM, July 2007, pp. 494–501.

[2] D. Cardillo, A. Rapp, S. Benini, L. Console, R. Simeoni, E. Guercio, and R. Leonardi, "The art of video MashUp: Supporting creative users with an innovative and smart application," *Multimedia Tools and Applications*, vol. 53, no. 1, pp. 1–23, May 2011.

[3] U.-N. Yoon, M.-D. Hong, and G.-S. Jo, "Automatic interactive video authoring method via object recognition," in *Intelligent Information and Database Systems —Ninth Asian Conference, ACIIDS 2017, Kanazawa, Japan, April 3–5, 2017, Proceedings, Part I*, ser. Lecture Notes in Artificial Intelligence, N. T. Nguyen, S. Tojo, L. M. Nguyen, and B. Trawiński, Eds., vol. 10192. Kanazawa, Ishikawa, Japan: Springer International Publishing, April 2017, pp. 589–598.

[4] Y. Liu, T. Mei, and C. W. Chen, "Automatic suggestion of presentation image for storytelling," in *Proceedings of 2015 IEEE International Conference on Multimedia and Expo*. Seattle, WA, USA: IEEE, July 2016, pp. 1–6.

[5] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "DBpedia —A crystallization point for the Web of data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, pp. 154–165, September 2009.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*. Lake Tahoe, NV, USA: Curran Associates, Inc., December 2013, pp. 3111–3119.

[7] B. L. Tseng, C.-Y. Lin, M. Naphade, A. Natsev, and J. R. Smith, "Normalized classifier fusion for semantic visual concept detection," in *Proceedings of 2003 International Conference on Image Processing*, vol. 2. Barcelona, Cataluna, Spain: IEEE, September 2003, pp. 535–538.

[8] S. K. Divvala, A. Farhadi, and C. Guestrin, "Learning everything about anything: Webly-supervised visual concept learning," in *Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: IEEE, June 2014, pp. 3270–3277.

[9] D. Comaniciu and P. Meer, "Mean shift analysis and applications," in *Proceedings of the Seventh International Conference on Computer Vision*, vol. 2. Toronto, ON, Canada: IEEE, September 1999, pp. 1197–1203.

[10] P. Pudil and J. Hovovičová, "Novel methods for subset selection with respect to problem knowledge," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 2, pp. 66–74, March 1998.

[11] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proceedings of the Thirteenth International Conference on Computer Vision*. Barcelona, Cataluna, Spain: IEEE, November 2011, pp. 2564–2571.

[12] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Transaction on Graphics*, vol. 35, no. 4, pp. 110:1–110:11, July 2016.

[13] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management: An International Journal*, vol. 24, no. 5, pp. 513–523, 1988.

[14] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh International Conference on Computer Vision*, vol. 2. Toronto, ON, Canada: IEEE, September 1999, pp. 1150–1157.

[15] P. F. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," in *Proceedings of the Twenty-fourth British Machine Vision Conference*, T. Burghardt, D. Damen, W. Mayol-Cuevas, and M. Mirmehdi, Eds. Bristol, England, UK: BMVA Press, September 2013, pp. 13.1–13.11.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA, USA: IEEE, June 2015, pp. 1–9.