A Seminar report
on

USAGE BASED TAG ENHANCEMENT OF IMAGES

Submitted in requirement for the course

TRAINING AND SEMINAR (CSN-499)
of Bachelor in Computer Science and Engineering

by

SAURABH VERMA
(Enrollment No. 13114057)

Project undertaken at

BIG DATA EXPERIENCE LAB
Adobe Systems
Bangalore

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE - 247667 (INDIA)

SUMMER, 2016

CONTENTS

I. DISCLAIMER

This document is strictly private, confidential and personal to its recipients and should not be copied, distributed or reproduced in whole or in part, nor passed to any third party. With respect to non-disclosure agreement, this document is intended to serve only as seminar report on internship project and many details are hidden.

II. ACKNOWLEDGEMENTS

I would love to grab this opportunity to express my deep appreciation and gratitude towards Balaji Vasan Srinivasan[1], my mentor for his constant support, guidance and feedback at all stages of the project. His vision, experience and zeal greatly drove the project. I don't think the project would have been successful without him.

I am highly indebted to Prof. Niloy Ganguly[2] from IIT Kharagpur for his quick insights and experienced ideas on the project. His comments and suggestions were invaluable to the project.

Having been in constant discussions with Lab members at various stages of project, I could not thank them enough for filling me with new and diverse ideas. I could understand why it is important to interact with other researchers while working there.

I am also highly grateful to Anandhavelu N, Ameena Khaleel and Smitha Sim for giving me a memorable internship experience, ranging from work to team to fun events. I congratulate them for running a very successful Summer Internship 2016-17.

[1] Email: balsrini@adobe.com

[2] His personal website http://www.facweb.iitkgp.ernet.in/ niloy/

III. ABSTRACT

Appropriate tagging of images is at the heart of efficient recommendation and retrieval and is commonly used with search engines and content management systems for indexing image content. However, existing technologies in image tagging focuses on what the content contains than how it is perceived. However, for a practical utilization of these tags, it is paramount to understand the usage of the image and incorporate them with the tags for better retrieval and recommendation. In places where the usage of the content is captured from the associated content, little is done to semantically connect the textual and visual content for a homogeneous tag representation. In this work we propose a system that analyzes the usage of an image and utilizes the information thus obtained to enhance the image tags.

IV. INTRODUCTION

Image retrieval by querying visual contents has been on the agenda of the database, information retrieval, multimedia, and computer vision communities for decades [1] [2]. Search engines like Baidu, Bing or Google perform reasonably well on this task, but crucially rely on textual cues that accompany an image: tags, caption, URL string, adjacent text etc. This fact motivated us to look for associated textual content based cues. In recent years, deep learning has led to a boost in the quality of visual object recognition in images with fine-grained object labels [3] [4]. Methods like LSDA [5] are trained on more than 15,000 classes of ImageNet [6] (which are mostly leaflevel synsets of WordNet [7]), and annotate newly seen images with class labels for bounding boxes of objects. Object labels, if recognized, make images easily retrievable for queries with these concepts. However, these labels come with uncertainty. For some images, there is much higher noise in its visual object labels; so querying by visual labels would not work here.

Problem: These limitations of text-based search, on one hand, and visual-object search, on the other hand, suggest combining the cues from text and vision for more effective retrieval. Although each side of this combined feature space is incomplete and noisy, the hope is that the combination can improve retrieval quality. Unfortunately, images that show more sophisticated scenes, or emotions evoked on the viewer are still out of reach. The search results would best be retrieved by queries with abstract words (e.g. ”environment friendly”) or activity words (e.g. ”traffic”) rather than words that directly correspond to visual objects (e.g. ”car” or ”bike”). So there is a vocabulary gap, or even concept mismatch, between what users want and express in queries and the visual and textual cues that come directly with an image. This is the key problem addressed in this work.

Challenges: The availabilty of a suitable dataset which is not domain specific, consisted of textual content and huge enough for our experiments was one of the most daunting initial challenges. Moreover, we needed to extract utility from a set of images which do not correspond to a particular domain. Since we were trying to enhance a given set of tags, we were essentially required to prove that our set of tags are actually better than any baseline out there. There wasn't any specific section of text relevant to the image, so we also had to check if the text entity is indeed relevant to the image or not. These facts establish the fact that the problem is hard.

Approach and Contribution: To bridge the concepts and vocabulary between user queries and image features, we propose an approach that harnesses commonsense knowledge (CSK). Recent advances in automatic knowledge acquisition have produced large collections of CSK: physical (e.g. color or shape) as well as abstract (e.g. abilities) properties of everyday objects (e.g. bike, bird, sofa, etc.), subclass and partwhole relations between objects, activities and

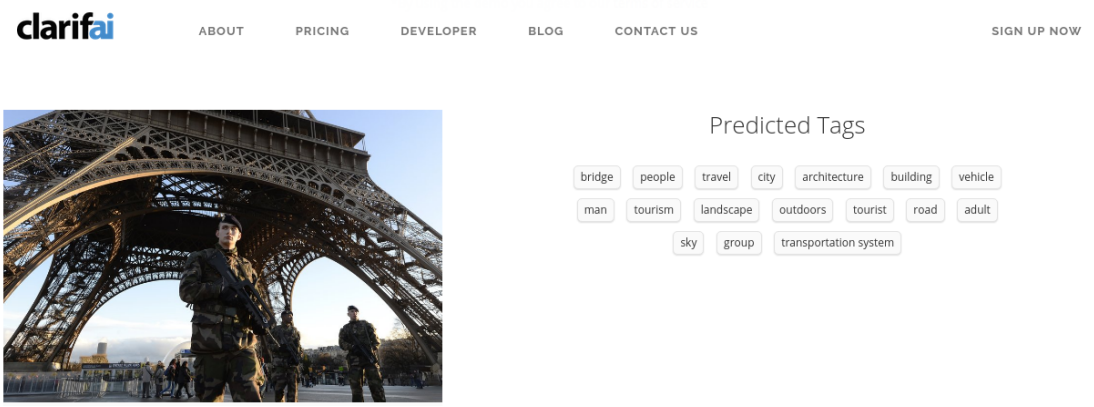


Fig. 1: Tags from Clarifai API

their participants, and more [8] [9] [10]. This kind of knowledge allows us to establish relationships between our entities and observable objects or activities in the image. This allows for retrieval of images with generic queries like ”big companies”. This idea is worked out into a query expansion model where we leverage a knowledge base for automatically expanding the entities with additional related entities. Our model unifies three kinds of features: textual features from the page context of an image, visual features obtained from recognizing fine-grained object classes in an image, and knowledge base features in the form of additional properties of the concepts referred to by entities. The weighing of the different features is crucial for optimal unification.

The papers contribution can be characterized as follows. We present the model for incorporating CSK into image retrieval. We develop a complete system architecture for this purpose. Our pipeline uses commonsense knowledge to enhance the set of tags for an image by looking at the components of the images in greater detail. We further discuss experiments that compare our approach to state-of-the-art auto tagging engines in various configurations. Our approach substantially improves the tag quality.

V. MOTIVATION

With the rise in popularity of social media, images accompanied by contextual text form a huge section of the web. Cisco estimates that annual global IP traffic will pass the zettabyte ([ZB] i.e. 1000 exabytes [EB]) threshold by the end of 2016, and will reach 2.3 ZB per year by 2020. Also, Global IP traffic will increase nearly threefold over the next 5 years.[3] This poses the need for efficient retrieval mechanisms, however, search and retrieval of documents are still largely dependent on solely textual cues. Although visual cues have started to gain focus, the imperfections in object/scene detection do not lead to significantly improved results. We also hypothesize that the use of background commonsense knowledge can significantly aid in retrieval of documents with associated images. Combining the above, we argue that an optimal combination of visual and textual features for retrieval along with application of common sense knowledge could significantly improve tags and hence, retrieval and recommendations. Similar motivation is argued for by Know2Look[11].

[3]The data is taken from Cisco Blog https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html

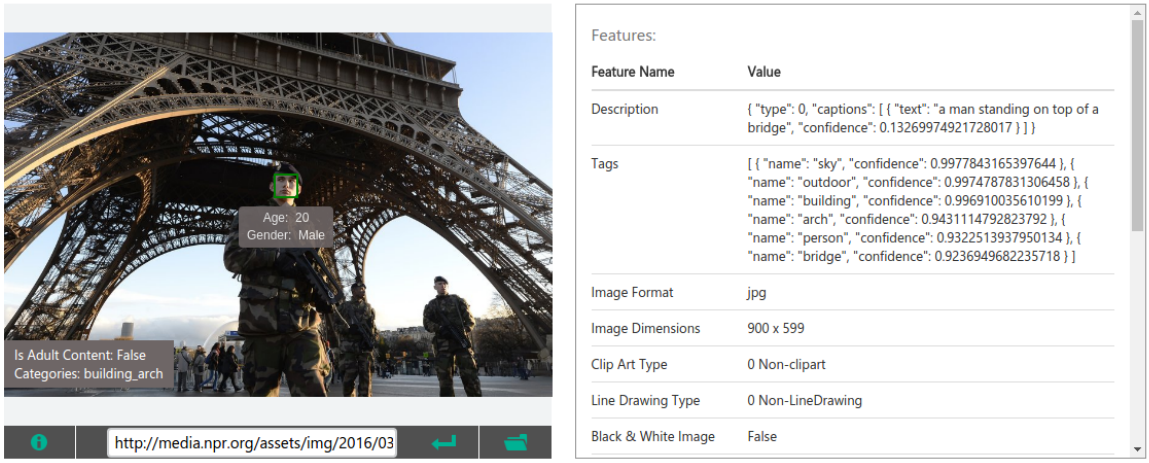


Fig. 2: Tags from Microsoft Vision API

Fig. 3: Tags from our pipeline

I would like to give an example here. The example image and report is taken from http://www.npr.org/sections/parallels/2016/03/22/471401729/in-vulnerable-europe-a-third-major-terrorist-attack-in The textual content around the image was following.

French soldiers patrol in front of the Eiffel Tower on Jan. 8, 2015, the day after a deadly attack on the French satirical newspaper Charlie Hebdo. With bombings Tuesday in Brussels, Europe has now been hit by three major terrorist attacks in just over a year. The terrorist attacks in Brussels mark the third major assault in the heart of Europe in just over a year and raise a troubling question: Are European states prepared to deal with a sustained onslaught? The satirical magazine Charlie Hebdo was hit in January 2015. Terrorists rampaged through Paris again in November. And now, Brussels has suffered bomb attacks at the airport and the subway, claiming more than 30 lives. ”This is not over,” French President Franois Hollande said just last Friday. He spoke in Brussels shortly after the arrest of Salah Abdeslam, believed to be the lone survivor of the Paris attack last fall. Hollande added that French and Belgian authorities had uncovered a much wider network of jihadists during their recent investigations. Yet this heightened state of alert did not prevent Tuesday’s bombings in Brussels. Analysts point to several factors behind the current attacks and the difficulty in stopping them. Paris and Brussels are major crossroads, with huge numbers of travelers crossing national boundaries,

making it difficult to track extremists. Prior to last year, Western Europe had not been hit hard by terrorists for a decade and therefore may have been underprepared. And the extremism so prevalent in the Middle East has now taken root in Europe. Map of Recent Attacks in Brussels and Paris The terrorist attacks in Brussels mark the third major assault in the heart of Europe in just over a year. Brussels and Paris are about 160 miles apart a little bit closer than the distance between New York City and Washington, D.C. In the map below, today's attacks in Brussels are marked in yellow; the November 2015 attacks in Paris in red; and the January 2015 attack at Charlie Hebdo in blue. "It's really only with the rise of ISIS in Syria and the attacks in Paris, and now this attack, that we've seen Western Europe facing such a concentrated, deadly and really sophisticated threat," Michael Leiter, the former director of the U.S. National Counterterrorism Center, told NPR's Morning Edition. "We're seeing the challenges of relatively open borders and a fractured intelligence system that makes it hard to detect and stop these attacks," he added. What's striking is that all three attacks were carried out in places that security experts considered potential targets. The Charlie Hebdo office was destroyed by a firebomb back in 2011, and after that, the magazine increased security as it carried on with its provocative cartoons that often lampooned Islam and Muslim figures. In the wake of the deadly 2015 attack on Charlie Hebdo, security was ramped up dramatically throughout Paris, including troops in the streets. Yet that didn't prevent the November attack at multiple sites that left 130 dead. And as French authorities investigated the Paris bloodbath, attention shifted to Brussels, where several attackers had been living and Abdeslam was captured last Friday. A woman visits a memorial Nov. 16 near the Bataclan concert hall in Paris. The attack on multiple locations in Paris last fall left 130 dead. A woman visits a memorial Nov. 16 near the Bataclan concert hall in Paris. The attack on multiple locations in Paris last fall left 130 dead. Analysts had warned that Western Europe was not immune. France and other European countries have large and often restive Muslim populations. Thousands of young Muslim men in Europe have traveled to and from Europe to join the Islamic State, according to various estimates. An estimated 200 have gone from Belgium, which gives it one of the highest per capita rates in Europe. "When I heard about the number of foreign fighters going in and out of Syria in 2013, I felt this wave would be coming to Europe and likely the United States," said Barry Pavel, director of the Brent Scowcroft Center at the Atlantic Council and a former member of the National Security Council under both President Obama and President Bush. "I'm not saying these attackers today came out of Syria. I don't know that. But as long as the Syria war continues, it's going to be an incubator for extremists," he added. Belgian officials hailed last Friday's arrest of Abdeslam as an important breakthrough. By taking him alive, they said, he could provide valuable information about the attack he's been linked to as well as other potential attacks. Yet it took authorities almost four months to find him in the neighborhood where he grew up, suggesting a strong network of sympathizers who protected him. Belgium's Foreign Minister Didier Reynders said Sunday that police had found heavy weapons last week in raids that led to Abdeslam's arrest. The suspect has been cooperating and told authorities that at least 30 jihadists remained at large in the city. "He was ready to restart something from Brussels," Reynders said. "That is maybe the reality, because we have found a lot of heavy weapons and we have seen a new network around him."

Figure 1 shows tags given by Clarifai API and Figure 2 shows tags given by Microsoft Vision API for the corresponding image only through visual analysis. Figure 3 shows tags given by our pipeline after analyzing both the visual content and the textual content. We find a sweet spot between the two where the set of tags is most relevant.

VI. OBJECTIVES

The project was required to fulfill the below objectives.

1) Combine the use of visual and textual features for tagging of images
2) Enhance the given set of tags using background knowledge from large knowledge bases
3) Convert the entities into most useful conventional form
4) Guide the system to give out a most useful set of tags
5) Use the final tags to improve retrieval and recommendation engines
6) Develop a cross-platform and re-entrant pipeline

VII. WORK PLAN

The project spanned for 12 weeks. The outline of the schedule is as follows.

1) Week 1 - 2 : This phase required us to brainstorm for a large number of ideas in a given space to completely explore the space. The ideas spanned from existing works to novel creations.
2) Week 3 : After having lots of ideas, we were required to finalise and present a problem alternative. We needed to establish it's business value and innovation value for the community.
3) Week 4 : Given a problem statement, this week needed us to design an appropriate solution framework which would not only solve the problem but also check for feasibility of the problem.
4) Week 5 - 8 : Having the design ready, we proceeded towards next steps in Software development Life cycle, namely, Coding and Testing. We quickly prototyped the high level ideas in solution framework, checked them and put them in place.
5) Week 9 - 10 : After building an end-to-end pipeline, we moved on the tricky part in our system, Evaluation, Tuning and Measurements. We had the responsibility to prove the utility of our tags with some concrete results. We tuned our pipeline and achieved outstanding results.
6) Week 11 : After finalizing on the pipeline and the results, this week required us to present our work to global research community of the organization in a very short span of time. We iterated over different presentations to give out the best we could and raised the bar for the coming presentations as we were first in.
7) Week 12 : Final week and my internship ended with Documentation and Code Transfer. I made everything run on their internal systems and improved the readability of my code, made improvements and signed off with a nice work experience.

VIII. DESCRIPTION

I would like to reemphasise the fact that most of the details have been intentionally hidden pertaining to Intellectual Property guidelines and Non-Disclosure Agreement.

A. Modules

The input to the system is an image with a set of associated content. First, that goes into text parser which outputs entities to Dependency Parser. After representing textual features, a set of author tags or image tags automatically extracted from visual tagging engines are

unified in the unifier module. After that, we use knowledge base to expand the set of entities in enhancement module. Finally, we extract most significant set of entities from the representation in Extraction Module.

B. Parsing

The input to the pipeline is an image and set of associated textual content. This module parses the text into entities, extracts out POS tags, named entity recognition, coreference resolution and all the other natural language processing tasks needed. Our system takes care of N-grams and different forms of entities in the text.

C. Representation

We needed a suitable representation to represent most of the information extracted from the text. There were different possibilities including first order logic, graphical representation, vector space representation and so on. All of them had their own pros and cons. We chose a particular representation to continue with all over our pipeline. This representation incorporated entities and their relationships.

D. Unifier

We also obtained a set of image tags or author tags from an external source for given image. Now we had to merge the same into the previous representation for entities extracted from text. This challenge was solved with proprietary algorithm which was our own novel work.

E. Enhancement

This module uses background knowledge base, concept networks and commonsense knowledge to enhance the set of entities with a richer set of entities and crafts the representation for final use. We iteratively expand the set of tags to a richer, diverse and relevant set. This is one of the unfamiliar and recent directions of popular works.

F. Extraction

Given a graphical representation of tags, we needed to convert them back to conventional form of key-value pair for use in end-use cases of recommendation and retrieval. We do the same by extracting top nodes from the graph using random walk based methods. Some of the readings included were [12] [13] [14]. We moved on with a very specific work from a tier-1 conference in the same domain and extended the same for our use.

IX. Implementation

A. Experimental Setup

We obtained gold standard human annotations for a dataset of 300 images. We also had access to multiple visual tagging engine APIs namely Clarifai, Imagga, Microsoft Vision API. We also implemented a text-based baseline[15] for comparison with extractive systems. For

comparison, we developed multiple metrics which measured utility, relevance and diversity of tags. After comparing them all, we established the fact that our tags are more relevant, diverse and significant than state-of-the-art systems.

B. Datasets

Conventional Datasets with no surrounding textual content like Corel[16], ESP-Game, IAPRTC-12 [17] cannot be used since our method extracts textual features as well. The dataset for the same was obtained from previous work by [15] which can be downloaded from http://lit.csci.unt.edu/index.php/Downloads.

C. Parameters

We needed multiple parameters for thresholds and importance ratios. Some of the significant parameters were as follows.

- A threhold to determine if two entities match.
- A threshold to create an edge in the graph while unifying visual cues.
- The ratio of probabilities with which we should move to the adjacent node in random walk.
- A threshold confidence in tags to include them in final output.
- A weighting scheme for normalization of tag weights from multiple sources.

D. Tools

There were multiple tools used at different stages to quickly develop a robust end-to-end pipeline. We used Stanford CoreNLP Natural Language Processing toolkit for the parsing and tagging functions[18]. We used entity disambiguation tool AIDA[19] to map entities in text. We used YAGO[8] as our knowledge base. Word2Vec[20] was used to represent words into vector space and work with them. We also used doc2vec[21] to convert document into vector-based representation. These are a very powerful set of tools provided by eminent research groups from Stanford University, Google and Max Planck Institute.

X. Results

We have achieved significant improvement in utility of tags with around 45% increase. Our tags are more relevant to the image in hand compared to baselines with an increase of 10%. Not to forget, we have great diversity in our set of tags. Measuring by one of the most standard measures, which has been popular since the 1960s, there is nearly an improvement of 25%. The improvements are measured over a dataset of 300 images and results are compared by using median of the achieved scores.

XI. Conclusions

With this paper, we propose use of both visual features and textual features along with background common sense knowledge for image retrieval. We focus on both connotational and denotational features of the image, i.e., how the image is percieved by the viewer and

what the image primarily contains. This gives us a balanced metadata(tags) to use in indexing which can be used to improve both retrieval and recommendation. Our pipelinr extracts visual and textual features, unifies them into a single representation and then enhances the entities derived to a more useful level with backround knowledge. By utilizing the combination of textual, visual and commonsense modalities we make the set of tags more appealing to the humans than traditional approaches. We support our claim by results of an internal survey conducted at the lab. Also, we prove relevance, significance and diversity of our set of tags objectively with multiple metrics and report an outstanding performance. The proposed concept could be easily extended for documents, videos or any multimedia.[15] is a text-based baseline.

REFERENCES

[1] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern recognition*, vol. 40, no. 1, pp. 262–282, 2007.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, p. 5, 2008.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] J. Hoffman, S. Guadarrama, E. S. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko, "Lsda: Large scale detection through adaptation," in *Advances in Neural Information Processing Systems*, pp. 3536–3544, 2014.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255, IEEE, 2009.

[7] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[8] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*, pp. 697–706, ACM, 2007.

[9] N. Tandon, G. de Melo, A. De, and G. Weikum, "Knowlywood: Mining activity knowledge from hollywood narratives," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 223–232, ACM, 2015.

[10] H. Liu and P. Singh, "Conceptnet — a practical commonsense reasoning tool-kit," *BT Technology Journal*, vol. 22, pp. 211–226, Oct. 2004.

[11] S. N. Chowdhury, N. Tandon, and G. Weikum, "Know2look: Commonsense knowledge for visual search,"

[12] P. Sarkar and A. W. Moore, *Random Walks in Social Networks and their Applications: A Survey*, pp. 43–77. Boston, MA: Springer US, 2011.

[13] K. P. Chitrapura and S. R. Kashyap, "Node ranking in labeled directed graphs," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pp. 597–606, ACM, 2004.

[14] S. Bhagat, G. Cormode, and S. Muthukrishnan, "Node classification in social networks," in *Social network data analytics*, pp. 115–148, Springer, 2011.

[15] C. W. Leong, R. Mihalcea, and S. Hassan, "Text mining for automatic image tagging," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 647–655, Association for Computational Linguistics, 2010.

[16] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *European conference on computer vision*, pp. 97–112, Springer, 2002.

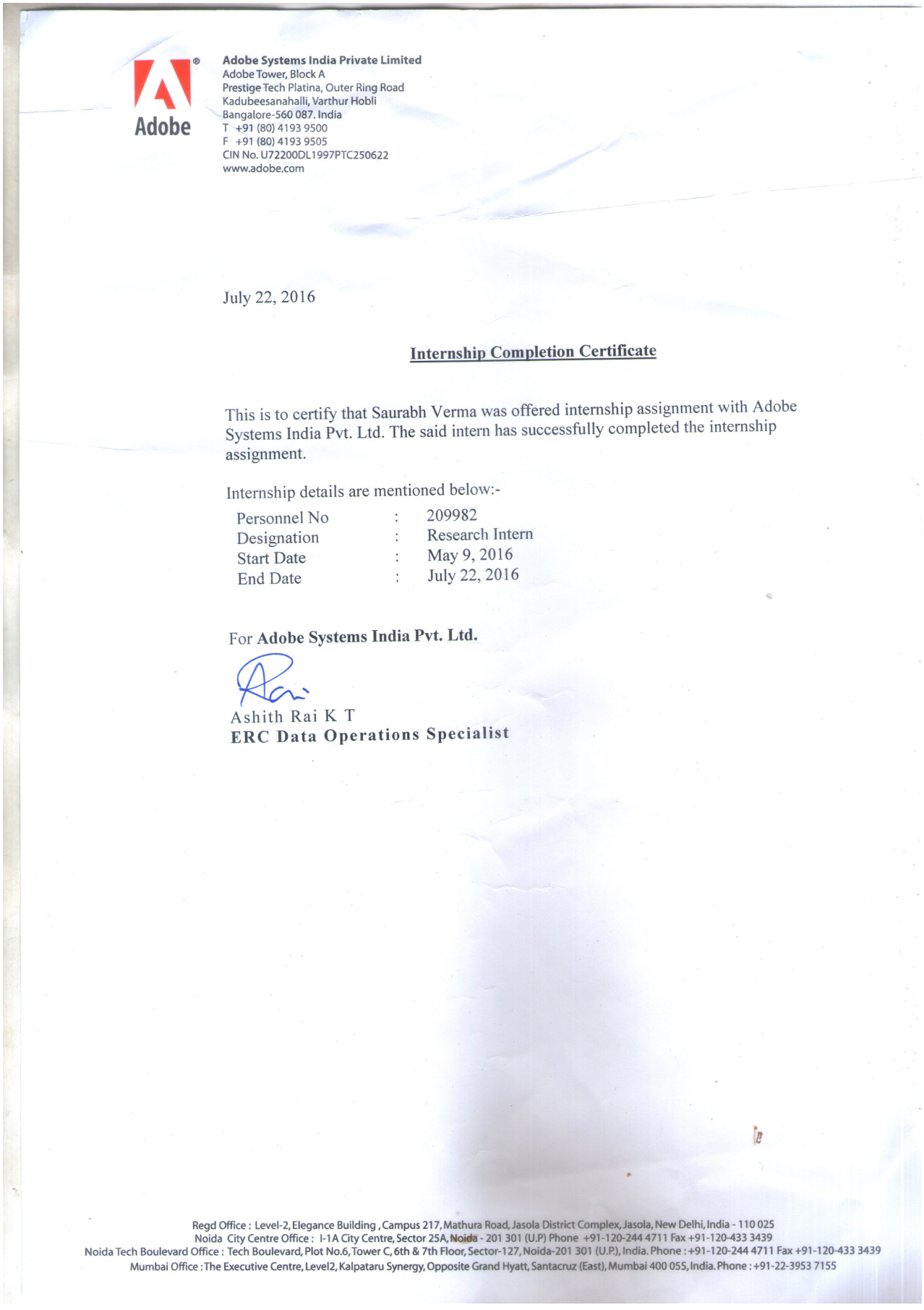
[17] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The iapr tc-12 benchmark: A new evaluation resource for visual information systems," in *International Workshop OntoImage*, vol. 5, p. 10, 2006.

[18] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit.," 2014.

[19] M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol, and G. Weikum, "Aida: An online tool for accurate disambiguation of named entities in text and tables," *Proceedings of the VLDB Endowment*, vol. 4, no. 12, pp. 1450–1453, 2011.

[20] T. Mikolov and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, 2013.

[21] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents.," in *ICML*, vol. 14, pp. 1188–1196, 2014.



Adobe Systems India Private Limited
Adobe Tower, Block A
Prestige Tech Platina, Outer Ring Road
Kadubeesanahalli, Varthur Hobli
Bangalore-560 087. India
T +91 (80) 4193 9500
F +91 (80) 4193 9505
CIN No. U72200DL1997PTC250622
www.adobe.com

July 22, 2016

Internship Completion Certificate

This is to certify that Saurabh Verma was offered internship assignment with Adobe Systems India Pvt. Ltd. The said intern has successfully completed the internship assignment.

Internship details are mentioned below:-

Personnel No	:	209982
Designation	:	Research Intern
Start Date	:	May 9, 2016
End Date	:	July 22, 2016

For **Adobe Systems India Pvt. Ltd.**

Ashith Rai K T
ERC Data Operations Specialist

Regd Office : Level-2, Elegance Building , Campus 217, Mathura Road, Jasola District Complex, Jasola, New Delhi, India - 110 025
Noida City Centre Office : I-1A City Centre, Sector 25A, Noida - 201 301 (U.P) Phone +91-120-244 4711 Fax +91-120-433 3439
Noida Tech Boulevard Office : Tech Boulevard, Plot No.6, Tower C, 6th & 7th Floor, Sector-127, Noida-201 301 (U.P.), India. Phone : +91-120-244 4711 Fax +91-120-433 3439
Mumbai Office : The Executive Centre, Level2, Kalpataru Synergy, Opposite Grand Hyatt, Santacruz (East), Mumbai 400 055, India. Phone : +91-22-3953 7155

Fig. 4: Internship Certificate from Adobe Systems