Image and Video Tag Aggregation

Olga Kanishcheva[0000-0002-4589-092X] and
Natalia Sharonova[0000-0002-8161-552X]

National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine
kanichshevaolga@gmail.com
nvsharonova@ukr.net

Abstract. In this paper, we explore the task of tag aggregations for the video and image files. In our work, we describe recent achievements image description generation, discuss tags and tagging. We present the result of our experiments based on the lexical resources and natural language processing approaches. In our work, we use the auto-tagging program from Imagga company as the generating program. As data, we use 5 videos which were split into shots for future processing. We identified two subtasks for tag aggregation: 1) creation of general set tags for the whole video file and 2) creation separate sets of tags for each video shots. Our detailed tag analysis and experiment results showed that pipeline with NLP methods received good results for the whole video, but for each shot, in future work, we need to use such resources as Nasari vectors, SensEmbed vectors or word2vec. We present all our experiments with graphs, tables etc. Our future work will be related to aggregation of video descriptions.

Keywords: Image Description · Video Description · Natural Language Processing · Aggregation of Video Tags · Evaluation Measures · Automatic Image Description Generation · Keywords Aggregation · Tag Aggregation

1 Introduction

Online social networks are providing more and more convenient services to their users. Today, social networks have grown to be one of the most important sources for people, they are involved in all aspects of their lives. Meanwhile, every online social network user is a contributor to such large amounts of information. Online users like to share their experiences and to express their opinions on virtually all events and issues. Among a large amount of online user-generated data, we are particularly interested in peoples opinions or sentiments towards specific topics and events.

In social web there appeared many aspects for exploration of large multimedia datasets that have previously been unavailable. Popular social websites, such as Flickr, Photobucket, Picasa etc. contain a massive amount of visual photographs, which have been collectively tagged and annotated by members of the respective community.

Other sources of images are different professional stock image marketplaces. Stock photos are made by professional or semi-professional photographers and are usually contained in search databases. They can be purchased and delivered online.

Each of these photos independent of a source should have relevant tags, sometimes these tags from human, sometimes this is auto-tags. The problems in the tags area are tag generation, tag translation/tag disambiguation, image classification/clusterization based on tags etc.

However, along with the growth in the number of images on the Internet, the number of video content growth also. The advances in computer and network infrastructure together with the fast evolution of multimedia data has resulted in the growth of attention to the digital videos development. The scientific community has increased the amount of research into new technologies, with a view to improving the digital video utilization: its archiving, indexing, accessibility, acquisition, store and even its processing and usability. Image and video processing are very close to each other and should be explored in parallel because the video can be divided into slots, where each slot represents an image.

The separate trend which is related to images and video is automatic generation of image description (full sentences). However, the problem of video description generation has several properties that make it especially difficult. Authors in the work [1] wrote that "Besides the significant amount of image information to analyze, videos may have a variable number of images and can be described with sentences of different length. Furthermore, the descriptions of videos use to be high-level summaries that not necessarily are expressed in terms of the objects, actions, and scenes observed in the images. There are many open research questions in this field requiring deep video understanding. Some of them are how to efficiently extract important elements from the images (e.g. objects, scenes, actions), to define the local (e.g. fine-grained motion) and global spatio-temporal information, determine the salient content worth to describe, and generate the final video description. All these specific questions need the attention of computer vision, machine translation and natural language understanding communities in order to be solve".

One of our goals is the construction of a system that optimizes the number of tags describing video resources, without any loss of sense. By using the textual information, a user is facilitated on the one hand to locate a specific video and on the other hand is able to comprehend rapidly the basic points.

The specificity of this problem lies in the fact that we have many tags from the auto-tagging program, and the keywords are necessary for the whole video and also for shots of these videos, because a user may need to find a fragment of some video. Therefore, this task is divided into two subtasks: 1) the optimization of keywords for the whole video file; 2) the aggregation of tag sets for separate shots.

The paper is organized as follows. In Section 2, we show the most recent related works for tag aggregation and in Section 3 we describe the Imaggas auto-tagging program that we use. The general approach, our methods for tag aggregation and results we analyze in Section 4. Section 5 concludes this paper.

2 Background and Related Work

The last three years have been associated with increased interest in the field of generating descriptions and keywords/tags for images. This task involved both large companies like Google, Microsoft and small, which are profiled on certain domain areas such as Clarifai (clarifai.com), Imagga (imagga.com) etc. Also in this area, preliminary studies and prerequisites for such intensive development have been made, for example, the emergence of special collections of images, such as ImageNet and Microsoft COCO etc. All this has allowed achieving that in 2014, research scientists on the Google Brain team trained a machine learning system to automatically produce captions that accurately describe images. Further development of that system led to its success in the Microsoft COCO 2015 image captioning challenge, a competition to compare the best algorithms for computing accurate image captions, where it tied for first place.

In the paper "Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge" [2], published in IEEE Transactions on Pattern Analysis and Machine Intelligence authors showed some successful examples. Today the program works with 94% accuracy, its a very good result.

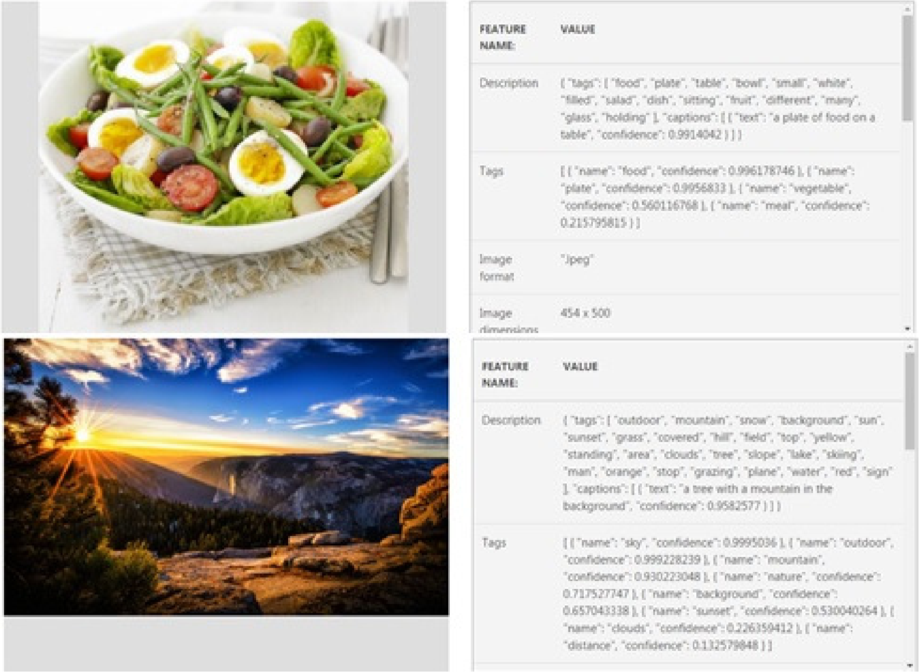
Another company Microsoft also has excellent results in this field. This company presents some results of image auto-tagging and image captioning. You can see that each image from Fig. 1 has Description field. This field consists of tag set and caption with the value of confidence. The Tags field also contains important tags with the relevant value of confidence.

In work [3], authors present a survey of models, datasets and evaluation measures for automatic description generation form images. In this paper presents approaches for tag-tagging as for description generation. For example, Kulkarni et al. System for labeling and generation sentences. They wrote that all models in this category achieve this using the following general pipeline architecture:

1. Computer vision techniques are applied to classify the scene type, to detect the objects present in the image, to predict their attributes and the relationships that hold between them, and to recognize the actions taking place.

2. This is followed by a generation phase that turns the detector outputs into words or phrases. These are then combined to produce a natural language description of the image, using techniques from natural language generation (e.g., templates, n-grams, grammar rules).

In paper [4] authors propose a tag-based framework that simulates human abstractors ability to select significant sentences based on key concepts in a sentence as well as the semantic relations between key concepts to create generic summaries of transcribed lecture videos. Their approach extractive summarization method uses tags (viewer- and author-assigned terms) as key concepts. They use Flickr tag clusters and WordNet synonyms to expand tags and detect the semantic relations between tags. This method could select sentences that have a greater number of semantically related key concepts.



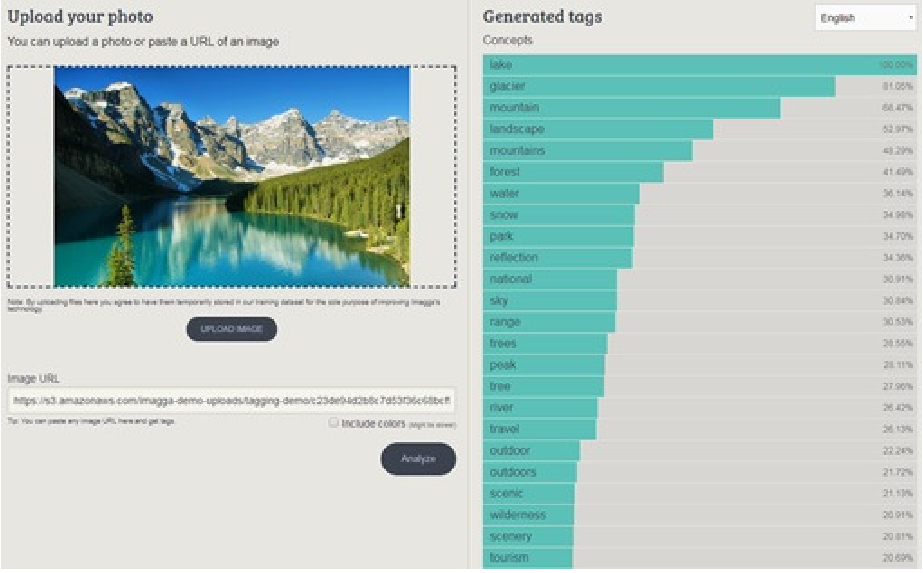
FEATURE NAME:	VALUE
Description	{ "tags": ["food", "plate", "table", "bowl", "small", "white", "filled", "salad", "dish", "sitting", "fruit", "different", "many", "glass", "holding"], "captions": [{ "text": "a plate of food on a table", "confidence": 0.9914042 }] }
Tags	[{ "name": "food", "confidence": 0.996178746 }, { "name": "plate", "confidence": 0.9956833 }, { "name": "vegetable", "confidence": 0.560116768 }, { "name": "meal", "confidence": 0.215795815 }]
Image format	"Jpeg"
Image dimensions	454 x 500

FEATURE NAME:	VALUE
Description	{ "tags": ["outdoor", "mountain", "snow", "background", "sun", "sunset", "grass", "covered", "hill", "field", "top", "yellow", "standing", "area", "clouds", "tree", "slope", "lake", "skiing", "man", "orange", "stop", "grazing", "plane", "water", "red", "sign"], "captions": [{ "text": "a tree with a mountain in the background", "confidence": 0.9582577 }] }
Tags	[{ "name": "sky", "confidence": 0.9995036 }, { "name": "outdoor", "confidence": 0.999228239 }, { "name": "mountain", "confidence": 0.930223048 }, { "name": "nature", "confidence": 0.717527747 }, { "name": "background", "confidence": 0.657043338 }, { "name": "sunset", "confidence": 0.530040264 }, { "name": "clouds", "confidence": 0.226359412 }, { "name": "distance", "confidence": 0.132579848 }]

Fig. 1. The examples from Microsoft (https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/).

3 Auto-Tagging of Images

The company IMAGGA (imagga.com) has developed an original technology for image auto-tagging by English keywords. The technology is based on machine learning and assigns to each image a set of keywords depending on shapes that are recognized in the image. For each learned item the system ”sees” in an image, appropriate tags are suggested. In addition, the system proposes more tags based on multiple models that it has learned. They relate the visual characteristics of each image with associated tags of similar images in ImageNet or big external manually created data sets (e.g. Flickr). The intuition and motivation are that more tags serve better in searching because users may express their requests by different wordforms. The platform developers believe they have found the right practical way to offer best possible image annotation solution for a lot of use-cases.



Upload your photo

You can upload a photo or paste a URL of an image

Note: By uploading files here you agree to have them temporarily stored in our training dataset for the sole purpose of improving Imagga's technology.

Image URL: https://s3.amazonaws.com/imagga-demo-uploads/tagging-demo/c23de94d2b6c7d53f36c68bcf

Tip: You can paste any image URL here and get tags.

○ Include colors (might be slower)

Generated tags

Concepts

lake	100.00%
glacier	81.05%
mountain	68.47%
landscape	52.97%
mountains	49.29%
forest	41.49%
water	36.14%
snow	34.98%
park	34.70%
reflection	34.36%
national	30.91%
sky	30.84%
range	30.53%
trees	28.55%
peak	28.11%
tree	27.96%
river	26.42%
travel	26.13%
outdoor	22.24%
outdoors	21.72%
scenic	21.13%
wilderness	20.91%
scenery	20.81%
tourism	20.69%

Fig. 2. Imagga's auto-tagging platform with automatically generated tags and their relevance scores: lake 100%, glacier 81,05%, mountain 68,47%, landscape 52,97%, forest 41,49%, water 36,14%, snow 34,98%.

Quite often, when the image contains a close-up object, Imagga's platform assigns correctly the most relevant tags to the central object (Fig. 2). In the right part of Fig. 2 keywords are ordered according to their relevance score. Associating external tags imports additional keywords in the annotation of Imagga's images.

4 Experiments and Results

We used five fragments of films for our experiments: Batmobile, FC Barcelona, Hunger Games, Meghan Trainor, Remi Gaillard. All these films were divided into shots. The structure of these files you can see on the Table 1. We received sets of tags for all videos with using auto-tagging program from Imagga company.

Table 1. Information about test data sets.

Name of film	Number of shots	Number of tags
"Batmobile"	24	1,524
"FC Barcelona"	57	1,570
"Hunger Games"	60	1,555
"Meghan Trainor"	154	6,161
"Remi Gaillard"	58	1,936

4.1 Preprocessing stage.

The authors in the paper [5] showed the approach for refinement of image tags. Such cleaning of tags is very necessary for images from social photo services. We used only part of these methods [5] for tag aggregation. First, we need to delete the duplicate tags, the next step it is necessary to tackle plural and inflected forms.

At the first stage, we remove duplicates. There are quite a lot of them from 68% to 92% of all video tags. Then we process phrases which are found among the keywords. For example, if we have tags such as jelly, fish and jelly fish, then we leave only jelly and fish tags, and the jelly fish tag remove from the tag set. This choice is based on the fact that single tags have a higher score, and therefore more relevant to the image. At the next step, we get delete of inflective forms, such as playing, played etc. For English, this can be done using the popular Porter stemmer. It is giving the opportunity for to remove the repeat words. At the end of this process, we delete the keywords which are characterizing the color (example, blue, red etc.). They occur quite often, but they are not needed for video tagging.

The effect from our refinement of image tags we will describe below. The Fig. 3 shows results for files Batmobile, FC Barcelona, Hunger Games, Meghan Trainor, Remi Gaillard.

The graphs show that the preliminary processing stage is sufficient to significantly reduce the number of tags that describe the whole video. However, we have each tag has a relevant score, which allowed us to evaluate its effect on the number of tags. In [6] it was investigated that tags that have a score more or equals 20 are the most significant for the image. We selected the keywords that score ranges from 20 to 100. Thus, we have 44 tags for the file "Batmobile", 32 "FC Barcelona", 70 "Hunger Games", 57 "Meghan Trainor", 68 "Remi Gaillard". These tags are presented in Tables 2, 3, 4, 5, 6.

We think that our results make it possible to consider this task to be performed at a fairly good level. But if the first task about the total number of tags for the whole video file is clear, then the solution to the task of tag aggregation for separate shots cannot be solved so simply, because we have not repetitions. Consider the results for the movie "Batmobile" (Fig. 4). The pre-processing stage showed that for the video "Batmobile" we have 1,524 tags for all shots, and after all filters, we received 1,399 tags, i.e. reduced only 8% by the total. Each color represents one shot and for the file "Batmobile" we have 24 shots.

We analyzed our results and decided that we can use only tags with the relevant score>20%, but we got the results shown in Fig. 5. This figure shows that we have 6 shots that do not have any tags with score> 20% at all. This is not good since all fragments must have keywords. Of the remaining 18 shots, 11 of them have less than 10 tags, which is also few. All this showed that in this case, we cannot use the score value as a defining characteristic. It also makes changes to the generation of a common set of tags for the whole video, since some shots will not be represented in the final set. The low score is primarily due to the features of the algorithm work of the program Imagga auto-tagging

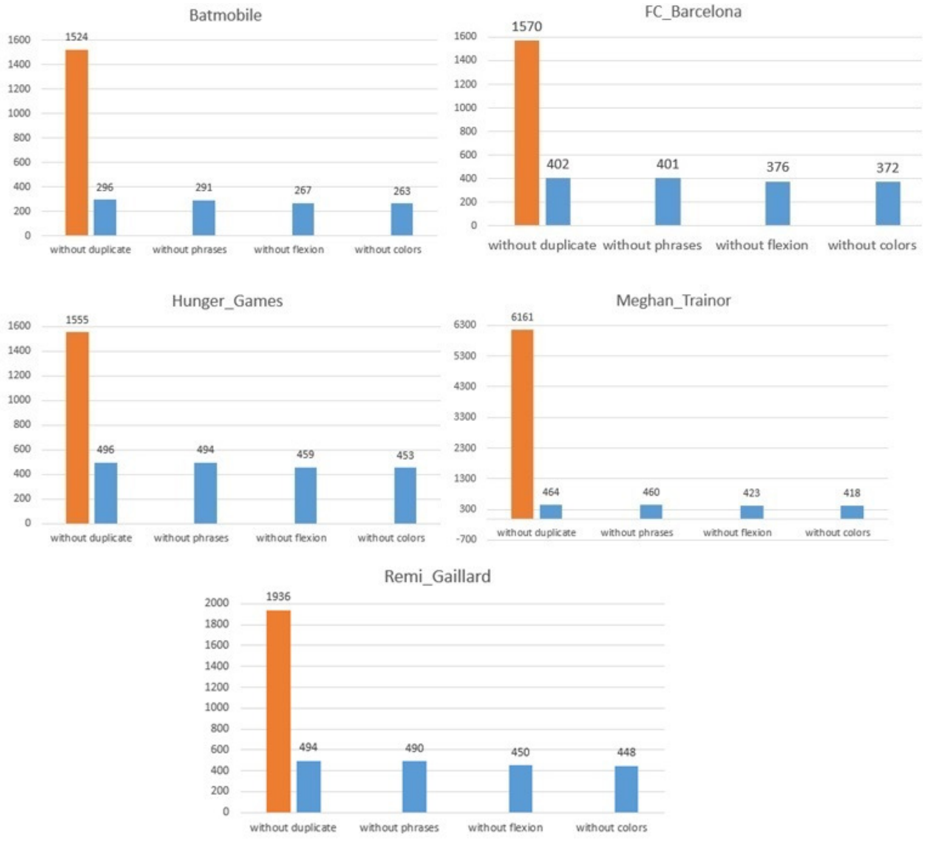


Fig. 3. Quantity of tags that are considered redundant after pre-processing stage.

Table 2. Set of tags with score >=20% for Batmobile file.

Tag	Score	Tag	Score	Tag	Score
car	100,00	businessman	32,00	team	23,53
man	40,94	convertible	31,88	wheel	23,39
container	39,63	automobile	31,82	sitting	23,17
petri dish	39,46	happy	29,57	meeting	22,93
vehicle	38,84	professional	28,92	job	22,57
dish	38,46	adult	28,75	success	22,22
people	37,02	smiling	28,14	attractive	21,86
business	35,38	person	27,89	women	21,66
auto	34,56	corporate	27,45	speed	21,64
businesswoman	34,22	work	26,39	tow truck	21,10
wheeled vehicle	33,65	transportation	26,21	drive	20,82
male	33,47	limousine	26,18	beaker	20,64
truck	33,40	crockery	24,66	suit	20,32
office	33,19	portrait	24,45	teamwork	20,06
caucasian	32,57	businesspeople	24,02		

Table 3. Set of tags with score >=20% for FC Barcelona file.

Tag	Score	Tag	Score	Tag	Score
swimsuit	70,54	caucasian	29,42	model	22,69
king	49,68	structure	28,36	summer	22,50
grass	44,63	people	27,00	happy	22,19
building	43,42	player	26,51	golfer	21,56
bikini	42,30	male	25,56	architecture	21,30
field	34,65	adult	24,84	torch	21,14
beach	34,32	person	24,79	body	21,13
man	33,45	attractive	24,60	ocean	20,79
greenhouse	32,91	smiling	24,00	silhouette	20,16
rival	32,71	sea	23,33	maillot	20,14
sexy	29,98	art	23,23		

Table 4. Set of tags with score >=20% for Hunger Games file.

Tag	Score	Tag	Score	Tag	Score
wheat	100,00	rural	29,18	hair	23,51
curtain	100,00	summer	29,00	adult	23,50
blind	95,77	water	28,72	landscape	23,43
furnishing	88,92	farm	28,06	icon	23,33
shower curtain	82,75	plant	27,66	straw	22,96
cereal	71,41	minaret	27,61	structure	22,85
protective covering	67,64	portrait	27,30	obstruction	22,53
fence	58,38	face	26,71	container	22,35
picket fence	56,33	blond	26,46	sign	22,28
field	45,64	seed	26,25	design	22,24
crystal	41,96	fountain	26,20	male	22,20
ice	40,13	building	26,10	river	21,71
covering	39,68	attractive	25,98	natural	21,53
menorah	39,44	source of illumination	25,31	solid	21,45
candle	38,91	candlestick	24,47	glass	21,43
dam	37,31	crop	24,45	sky	21,40
grain	37,13	man	24,25	symbol	21,25
barrier	34,24	person	24,20	model	21,15
clock	34,16	corn	24,18	coral fungus	21,14
photograph	32,18	people	24,11	caucasian	21,04
candelabrum	31,56	timepiece	23,97	bread	20,59
agriculture	30,35	pretty	23,74	looking	20,21
harvest	29,87	cleaning implement	23,70		
mosquito net	29,86	suit	23,64		

Table 5. Set of tags with score >=20% for Meghan Trainor file.

Tag	Score	Tag	Score	Tag	Score
art	63,63	plaything	28,47	blind	23,68
blond	56,92	happy	28,39	cute	23,47
candle	50,18	smile	28,37	body	22,93
dress	47,26	makeup	28,35	quill	22,58
toiletry	46,45	shower curtain	27,80	lady	22,44
nipple	41,64	gymnastics	27,71	eyes	22,31
attractive	36,84	pajama	27,70	nightwear	22,15
hair	34,83	source of illumination	27,41	clothing	21,98
portrait	33,98	gown	27,03	letter opener	21,68
face	32,89	silhouette	26,91	light	21,42
cap	32,28	sexy	26,32	net	20,74
model	32,22	hair spray	25,90	expression	20,52
pretty	31,89	swing	25,19	man	20,51
adult	31,36	complexion	24,88	cheerful	20,45
curtain	30,82	fashion	24,79	mechanical device	20,40
person	30,61	lipstick	24,13	sunset	20,30
people	29,73	top	24,05	human	20,19
caucasian	29,06	swimsuit	23,99	lips	20,05
texture	28,54	furnishing	23,71	pen	20,03

Table 6. Set of tags with score >=20% for Remi Gaillard file.

Tag	Score	Tag	Score	Tag	Score
curtain	87,58	maillot	34,72	sexy	23,96
shower curtain	75,19	seaside	33,90	coastline	23,79
swimsuit	73,50	truck	33,24	van	23,76
beach	68,46	coast	31,06	outdoors	23,75
furnishing	68,17	shepherd dog	31,03	grass	23,47
blind	67,44	sky	30,84	structure	23,18
shore	58,56	pool	30,50	minibus	22,33
swimming	47,97	fountain	29,09	adult	22,30
sea	47,63	swimming trunks	29,03	sport	22,27
protective covering	47,01	walk	29,00	road	21,93
vessel	46,64	landscape	28,85	vehicle	21,68
ship	45,68	summer	28,41	jump	21,49
car	44,25	vacation	28,33	art	21,45
boat	44,13	fireboat	28,28	ball	21,30
sand	42,51	german shepherd	28,23	swing	20,98
dune	42,10	travel	27,98	bus	20,92
bikini	41,88	dog	27,50	person	20,92
screen	40,04	male	27,09	minivan	20,92
ocean	39,66	man	26,95	fun	20,70
golf	39,63	covering	26,03	holiday	20,50
garment	37,09	surfing	25,51	people	20,36
water	36,47	sun	24,26	seashore	20,10
door	35,92	clothing	24,08		

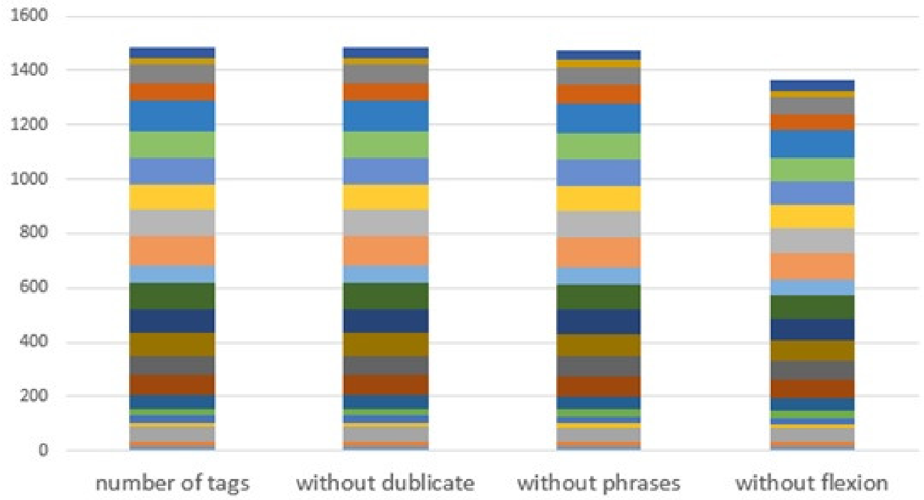


Fig. 4. Quantity of tags that are considered redundant after preprocessing stage of each shot for Batmobile film.

program. But it is getting better every year and so we think that in the future all shots will have tags with a high score.

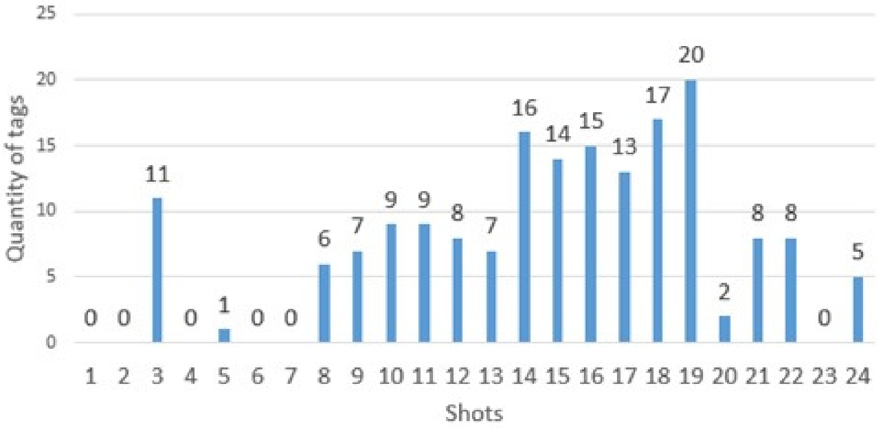


Fig. 5. Quantity of tags for each shot for the Batmobile film.

In future, we decided to make some experiments with Nasari vectors, SensEmbed vectors or word2vec for the aggregation of keywords for a single shot.

5 Conclusion

In this work we i) analyzed of systems which generate tags/descriptions for video and images; ii) compared the results of related works; iii) explored methods for generating natural language descriptions for images/video; iv) created short overview of the generation of video and image descriptions and explored main problems of this task. We concentrated on the problem of tag (keywords) aggregation into a single description of the object. Multimedia collections integrate electronic text, graphics, images, sound, and video. Their objects are usually annotated by keywords which characterize, describe or refer to categories in certain classifications. These tags help to distinguish the objects and often form folksonomies: user-generated categories for organizing digital content. In this work, we showed how works the preprocessing stage for tag optimization of keywords sets for video fragments, using NLP techniques, lexical resources to tag aggregation. We presented the statistical information about our experiments and results.

References

1. Peris A., Bolaos M., Radeva P., Casacuberta F.: Video Description Using Bidirectional Recurrent Neural Networks. In: Proceedings of the International Conference

on Artificial Neural Networks ICANN 2016: Artificial Neural Networks and Machine Learning, pp. 3-11, Barcelona, Spain (2016). https://doi.org/10.10007/978-3-319-44781-0.1

2. Vinyals O., Toshev A., Bengio S., Erhan D. Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 4, pp. 652-663 (2017).
3. Kulkarni G., Premraj V., Dhar S., Li S., Choi Y., Berg A. C., Berg T. L. Baby talk: Understanding and generating simple image descriptions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2891-2903 (2011).
4. Hyun Hee Kim, Yong Ho Kim. Generic Speech Summarization of Transcribed Lecture Videos: Using Tags and Their Semantic Relations. Journal of the Association for Information Science and Technology **67**(4), 366–379 (2016)
5. Kanishcheva O., Angelova G. A Pipeline Approach to Image Auto-Tagging Refinement. In: Proceedings of the 7th Balkan Conference on Informatics Conference, Craiova, Romania, ACM New York, NY, USA (2015) https://doi.org/10.1145/2801081.2801108
6. Kanishcheva O., Angelova G. About Emotion Identification in Visual Sentiment Analysis. In: Proceedings of the 10th International Conference on "Recent Advances in Natural Language Processing" RANLP 2015, , pp. 258-265, Hissar, Bulgaria (2015)