Weekly Progress Report

Worklet ID: 23RSG40SRM
Worklet Title: Filter Duplicate Images

Mentors: Abhishek Mishra, Ankit Mishra, Athira Menon
Faculty: Dr. Gayathri M, Dr. M. Suganiya
Students: Kanupriya Johari, Diptayan Jash, Tuhina Tripathi, Avya Rathod
College: SRM Institute of Science and Technology

Agenda:

Sr. No.	Description	Remarks	Progress
1	Literature survey and study	Thoroughly studied research papers concerning implemented methodologies.	Completed
2	Evaluate various image duplication checking and filtering methods	Evaluated and implemented p-hash, d-hash and Open-AI CLIP model.	Awaiting Approval
3	Write a custom script to find and filter out duplicates in images.	Written Custom scripts for p-hash, d-hash and CLIP model, SIFT algorithm.	Awaiting Approval
4	Test it for high scale and maintain accuracy.	-	Pending
5	Improve the algorithm to improve the decided parameters.	-	Pending
6	Write Research Paper stating innovative methods to find and filter duplication	-	Pending

Dataset Description:

Sr. No.	Dataset Name	Dataset Description	Link	Password
1	California-ND	Total Images - 701 Resolution: 1024x768 Authentic user-generated photos	Link	QoMEX-2013

Weekly Project Progress:

Sr. No.	Week	Date	Work Done	Remarks
1	**Week 1**	**16.01.2024**	Conducted a thorough literature survey, studied research papers concerning implemented methodologies such as various Hashing Techniques, Neural Networks etc.	Presented our findings to the mentors on 16.01.2024
2	**Week 2**	**23.01.2024**	Explored options for implementation.	Explored Hashing Techniques thoroughly. (p-hash, d-hash, w-hash, b-hash)
3	**Week 3**	**30.01.2024**	Conducted extensive dataset search; initiated communication with NUS professors for dataset acquisition.	Dataset - California-ND
4	**Week 4**	**06.02.2024**	Analyzed acquired dataset and successfully applied p-hash technique. Uploaded the script on GitHub.	In p-hash it takes around 35 seconds for completion, however we do find the duplicates in the dataset. Thus, not that efficient.
5	**Week 5**	**13.02.2024**	Implemented d-hash technique. Uploaded the script on GitHub.	We found out that d-hash was only working well for the exact duplicates.
6	**Week 6**	**20.02.2024**	Implemented CLIP: Connecting text and images model by OpenAI.	CLIP model takes approximately 3 minutes to run however it efficiently removes the near-duplicates and duplicate images.
7	**Week 7**	**27.02.2024**	Refining accuracy for our implementations. Tried to look for more methods.	Worked on ways to improve time complexity and refine our scripts even further.